Automatic Assessment of Speech Impairment in Cantonese-Speaking People with Aphasia

Ying Qin, *Student Member, IEEE*, Tan Lee, *Member, IEEE*, and Anthony Pak Hin Kong

***Abstract*—Aphasia is a common type of acquired language impairment resulting from dysfunction in specific brain regions. Analysis of narrative spontaneous speech, e.g., story-telling, is an essential component of standardized clinical assessment on people with aphasia (PWA). Subjective assessment by trained speech-language pathologists (SLP) have many limitations in efficiency, effectiveness and practicality. This article describes a fully automated system for speech assessment of Cantonese-speaking PWA. A deep neural network (DNN) based automatic speech recognition (ASR) system is developed for aphasic speech by multi-task training with both in-domain and out-of-domain speech data. Story-level embedding and siamese network are applied to derive robust text features, which can be used to quantify the difference between aphasic speech and unimpaired one. The proposed text features are combined with conventional acoustic features to cover different aspects of speech and language impairment in PWA. Experimental results show a high correlation between predicted scores and subject assessment scores. The best correlation value achieved with ASR-generated transcription is .827, as compared with .844 achieved with manual transcription. The siamese network significantly outperforms story-level embedding in generating text features for automatic assessment.**

***Index Terms*—Pathological speech assessment, aphasia, Cantonese, automatic speech recognition, deep neural network (DNN).**

I. Introduction

APHASIA is a common type of acquired language impairment resulting from dysfunction in specific brain regions. It is typically related to a stroke or other physical conditions such as head trauma or tumor. Aphasia may impair a person's ability to comprehend or formulate language, which could affect the communication modalities of auditory comprehension, verbal expression, reading, and writing [1]. People with Aphasia (PWA) show various types of symptoms, e.g., inability to pronounce, difficulty in naming objects and forming words, and/or comprehending language [2]. The symptoms depend on the location of injured brain region and vary in the degree of severity [3]. Their presence may have significant negative impact on daily communications of PWA, and lead to low self-esteem and social isolation [4].

Analysis of narrative spontaneous speech (e.g., picture description, story-telling) produced by PWA is an essential component of clinical assessment process for evaluating the severity and/or type of aphasia. The content and fluency of unprepared narrative speech are considered informative indicators of the severity of disorder [5], [6]. The assessment is carried out by trained speech-language pathologists (SLP) with pertinent linguistic and cultural background. The reliability and accuracy of such subject assessment approach depend on the clinician's experience. Its efficiency in practical use is limited by the need for manual transcription of speech, which is very time-consuming [7]. Due to the global shortage of SLPs, many PWA do not have the chance of being timely assessed and regularly monitored on the state of impairment. Computer-assisted assessment based on advanced signal processing and machine learning techniques is believed to be an effective means to address this problem.

Automatic analysis of aphasic speech has been investigated to assist diagnosis, treatment and rehabilitation of PWA. In earlier studies, a variety of text and acoustic features were derived from manually produced speech transcription and annotation to distinguish aphasic speech from normal one [8], [9]. With the advancement of automatic speech recognition (ASR) technology, fully automated approaches are actively explored in recent years [10], [11]. Feature extraction is performed based on text output and time alignment information generated by ASR systems. The feature design relies largely on expert knowledge acquired in clinical practices. Specifically, text statistics, e.g., word frequency count by part-of-speech, and psycho-linguistic knowledge, e.g., word-level familiarity and age of acquisition scores [12], [13], were shown to be useful indicators of language impairment.

Needless to say, in the ASR-based approach, the quality of ASR output plays a critical role in achieving reliable speech assessment. General-purpose ASR systems could not be straightforwardly applied to impaired speech [10], [14]. The mismatches in voice, articulation and language usage may lead to a high word error rate. Developing an application-specific ASR system with high accuracy is also a difficult task because of the scarcity

Manuscript received May 15, 2019; revised August 16, 2019; accepted November 8, 2019. Date of publication November 28, 2019; date of current version April 8, 2020. The work of T. Lee was supported in part by a GRF Project Grant CUHK14227216 from the Hong Kong Research Grants Council, in part by a Direct Grant from the CUHK Research Committee, in part by the CUHK Research Sustainability Fund, and in part by the CUHK Shenzhen Research Institute. The work of A. P. H. Kong was supported by the National Institutes of Health under Project NIH-R01-DC010398. The guest editor coordinating the review of this paper and approving it for publication was Prof. Douglas O'Shaughnessy. *(Corresponding author: Ying Qin.)*

Y. Qin and T. Lee are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (e-mail: yingqin@link.cuhk.edu.hk; tanlee@cuhk.edu.hk).

A. P. H. Kong is with the School of Communication Sciences and Disorders, University of Central Florida, Orlando, FL 32816 USA (e-mail: antkong@ucf.edu).

Digital Object Identifier 10.1109/JSTSP.2019.2956371

of disease-matched training data. Therefore, it is important to design clinically-relevant features that are robust to errors in ASR output [11].

The present research aims to develop a fully automated speech assessment system for Cantonese-speaking PWA. The system takes in narrative speech produced by the subject being assessed and makes prediction on the severity of aphasia based on the characteristics of input speech. The proposed system design is centered around a range of machine learning techniques that are applied to extract, model and classify impairment-related linguistic and acoustic features. First, a deep neural network (DNN) based ASR system is trained by multi-task learning strategy using a limited amount of in-domain data, supplemented by a large amount of out-of-domain data. Second, a word embedding model is applied to derive robust text features from speech transcription. Third, a siamese network is trained to contrast the topic relevance between a pair of narration transcripts. In addition, the newly designed text features are combined with conventional acoustic features to cover different dimensions of speech impairment. To the best of our knowledge, this is the most complete attempt to aphasic speech assessment for non-Western languages.

In the next section, previous research related to aphasic speech assessment is briefly reviewed. The Cantonese database of aphasic speech and the overall system design are described in Section III and IV respectively. The development of a dedicated ASR system is elaborated in Section V. Section VI and VII give details of the proposed text features and acoustic features, respectively. Experimental setup and results are presented in Section VIII and IX, respectively.

II. Related Work

Conventionally automatic assessment of pathological speech is formulated as a pattern recognition problem that involves an explicit process of feature extraction. Atypicalities in PWA are manifested in the aspects of dysfluency, shortage of vocabulary, paraphasia, etc. Fraser *et al.* investigated a set of text and acoustic features from narrative speech on classification of sub-types of primary progressive aphasia (PPA) [8], [9]. Extraction of the text features required manual transcription of speech, and thus the system did not support fully automated assessment. In [10], an off-the-shelf ASR system was utilized to generate text transcription for feature extraction. The ASR accuracy on impaired speech was found to be inadequate for reliable assessment in practice. The study by Peintner *et al.* [15] dealt with the problem of sub-type classification of Fronto-temporal Lobar Degeneration. The proposed features, including phone duration, pause duration, part-of-speech (POS), and word frequencies, were computed from text output and time alignment information produced by a general-purpose meeting transcription system.

Duc Le *et al.* proposed to use a comprehensive set of acoustic and linguistic features to predict subjective assessment scores on aphasic speech [11]. Extracted with the assistance of a tailor-made ASR system, the features were intended to measure information intensity, dysfluency, lexical diversity, and deviations in rhythm and pronunciation (with respect to normal speech). In order to make the features more robust to ASR errors, a linear transformation was applied to map the raw features derived from ASR output to calibrated ones as if they were computed based on the respective ground-truth transcription.

In recent years, new deep learning models have been introduced to integrate feature extraction as part of model training, and let feature design be data-driven. In [16], a long-short-term memory (LSTM) model trained with word embeddings was used for classifying aphasia types in Germany-speaking PWA. The classification accuracy was on the low side mainly because of the shortage of training data. In [17], an end-to-end "utterance-to-score" approach was attempted in aphasic speech assessment. While effective binary classification of severe and mild cases could be achieved, it is difficult to extend to more complicated problem unless the amount of training data can be significantly increased.

Lacking in-domain training data has been a major obstacle to the development of ASR system for handling disordered speech. Various methods were proposed to mitigate this problem. They include speaker adaptation [18], tandem-based feature extraction [19], speaker-specific pronunciations learning [20] and multi-task learning strategy [21]. It was also suggested to exploit out-of-domain or in-domain healthy speech in ASR system training [14], [22]. Duc Le *et al.* made multiple attempts to improving ASR acoustic model via adaptation and pre-training based on out-of-domain impaired speech databases [23], [24], and the multi-task learning approach [11]. Multi-task learning was also applied to improve the performance of a Cantonese ASR system for aphasic speech assessment [25]. The system adopts the structure of time-delay neural network combined with BLSTM (TDNN-BLSTM). The two auxiliary learning tasks involved two out-of-domain unimpaired speech databases and were shown to benefit the main task of topic-specific narrative speech from PWA.

III. Database: Cantonese AphasiaBank

Cantonese is a major Chinese dialect spoken by tens of millions of people in the provinces of Guangdong and Guangxi of Mainland China, Hong Kong, Macau, and overseas Chinese Communities. The Cantonese AphasiaBank (CanAB) is a large-scale multi-modal corpus developed jointly by University of Central Florida and the University of Hong Kong [26]. Its primary goal was to support both fundamental and clinical research on Cantonese-speaking aphasia population. The corpus contains audio recordings of narrative speech from 105 PWA and 149 unimpaired subjects. All of them are native speakers of Cantonese. The speech recordings were elicited following the English AphasiaBank protocol [27], [28], with adaptation to the local culture. The 9 narrative tasks are described in detail as in Table I. Each PWA was required to complete all of the 9 tasks, while each unimpaired subject was arranged to complete 8 tasks (since the "Stroke" task is not applicable). Except for personal monologue, each of the remaining 7 tasks is about specific topics, which is referred as a "story". A head-worn condenser microphone and a digital recorder were used for audio recording at sampling rate of 44.1 kHz.

TABLE I
DESCRIPTION OF 9 TASKS IN CANTONESE APHASIABANK

Category	Task	Description
Single picture description	CatRe	Black and white drawing of a cat on a tree being rescued.
	Flood	A color photo showing a fireman rescuing a girl.
Sequential picture description	BroWn	Black and white drawing of a boy accidentally breaking a window.
	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 telling	CryWf	Telling a story from a picture book "The boy who cried wolf".
	TorHa	Telling a story from a picture book "The tortoise and the hare".
Personal monologue	ImpEv	Description of an important event in life.
	Stroke	Description of the experience of suffering a stroke.

TABLE II
TWO GROUPS OF SPEAKERS FROM CANTONESE APHASIABANK USED IN THIS PAPER, NAMELY APHASIC SPEAKERS (*APHASIA*) AND UNIMPAIRED CONTROL SPEAKERS (*CONTROL*)

	Aphasia	Control
# of speakers	92	118
Age	54 ± 9	48 ± 16
Gender	60 Male, 32 Female	46 Male, 72 Female
Duration	16.4 hours	14.4 hours

The speech recordings were transcribed by trained research assistants using the CLAN (Child Language Analysis) program [29]. The orthographic transcription is in the form of a sequence of Chinese characters. Fillers, unintelligible speech and non-speech sounds are represented by special symbols as specified in CLAN. The Chinese characters are then converted into syllable transcription in the Jyut Ping format using a Cantonese pronunciation lexicon [30].

All impaired subjects in Cantonese AphasiaBank participated in a standardized comprehensive assessment using the Cantonese Aphasia Battery [6]. The assessment comprises a series of sub-tests measuring speech fluency, naming abilities, etc. The sum of sub-test scores is commonly termed as the Aphasia Quotient (AQ). The value of AQ ranges from 0 to 100, indicating the overall severity of impairment. Lower AQ value means higher degree of severity.

As shown in Table II, a total of 16.4 hours of speech data from 92 impaired subjects and 14.4 hours of speech data from 118 unimpaired subjects are available for use in this study.[1] They are named as the *Aphasia* group and *Control* group, respectively. The *Aphasia* group of subjects include 59 Anomic aphasia, 6 Transcortical sensory aphasia, 12 Transcortical motor aphasia, 10 Broca's aphasia, 1 Isolation aphasia, 2 Wernicke's aphasia and 2 Global aphasia. Their subjective AQ scores vary in the range of 11.0 to 99.0. Fig. 1 shows the histogram of AQ and Fig. 2 shows the distributions of AQ for 7 types of aphasia. Subjects diagnosed as Global aphasia obtain the lowest AQ scores. Subjects with high AQ scores are typically from the Anomic population.

[1]The other 13 impaired subject and 31 unimpaired subjects were not used in the study because the audio recordings and/or transcriptions are not complete for various non-technical reasons.

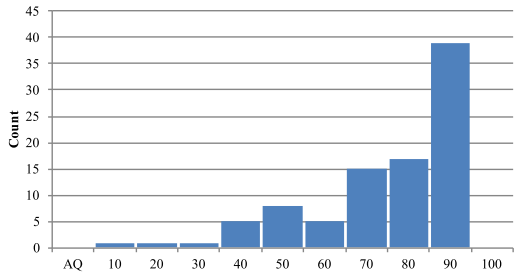


Fig. 1. Histogram of AQ scores of the 92 impaired subjects in the Cantonese AphasiaBank.

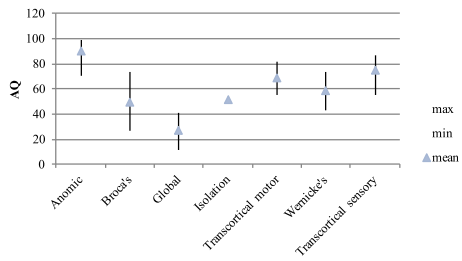


Fig. 2. Distributions of AQ values for each of the 7 aphasia types among the 92 impaired subjects.

IV. OVERALL SYSTEM DESIGN

As suggested in previous literature, speech impairment in PWA can be analyzed in linguistic aspect and acoustic aspect. The former one is based on text-based features, reflecting vocabulary and content related impairment. The latter one is reflected on the acoustic signal, e.g., dysfluency, voice changes, which can be derived from speech recordings. In order to achieve a more complete assessment, both aspects are covered in the proposed system design.

The overall architecture of the proposed system is shown as in Fig. 3. To begin with, the input speech from an impaired subject is decoded by an ASR system that is developed specifically to handle aphasic speech. The ASR system is trained with domain-matched data from unimpaired speakers in CanAB, as well as a few domain-mismatched Cantonese speech databases using the multi-task learning strategy. The ASR system can be configured to generate different types of representations of input speech, e.g., 1-best ASR output and confusion network. It can also generate time alignment information at phone level or above. In parallel to ASR, pitch and formant frequencies are

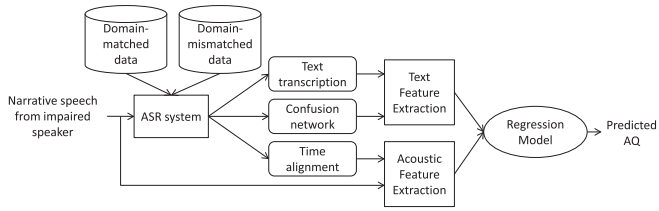


Fig. 3. Overall architecture of the proposed assessment system for aphasic speech.

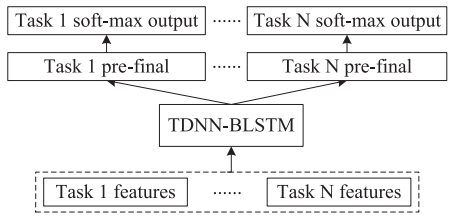


Fig. 4. Structure of multi-task TDNN-BLSTM acoustic model.

computed directly from the raw speech signal. The most critical modules of the system are text feature extraction and acoustic feature extraction. The extracted features are used to train a regression model and predict AQ score of a test speaker.

V. ASR System for Aphasia Assessment

A. Training and Test Data

The speech data used for acoustic model training are divided into two parts: in-domain data and out-of-domain data. Both parts are from unimpaired speakers.

1) In-Domain Data: There are a total 118 control subjects in the CanAB, each having completed 8 narrative tasks. The in-domain training data comprises about 12.6 hours of recordings from 101 unimpaired speakers. The remaining 1.8 hours of speech from the other 17 unimpaired subjects are used for ASR performance evaluation.

2) Out-of-Domain Data: King-ASR-086 & CUSENT: The out-of-domain data come from two publicly available Cantonese speech databases, both of which were created for large-vocabulary acoustic modeling.

- **King-ASR-086**
  King-ASR-086 (K086) [31] is a commercial speech corpus of Cantonese. It contains 87.4 hours of speech from 136 Cantonese speakers. The speech content is mainly news on a wide range of topics. In this study, 32,264 utterances from 55 male and 55 female speakers are used to train the ASR system for aphasic speech.
- **CUSENT**
  CUSENT is a large-scale continuous speech database of Cantonese developed by the Chinese University of Hong Kong [32]. The speech content includes 5,100 sentences selected from local newspaper articles, which provide a balanced phonetic coverage. The CUSENT training data contains 20,378 utterances from 34 male and 34 female speakers. The total duration is 19.3 hours.

The test data used for ASR performance evaluation are from the 92 PWA and the remaining 17 control subjects in the CanAB. Audio data from different databases are re-sampled to 16 kHz, regardless of the original sampling rates.

B. Acoustic Modeling

As described in [25], the multi-task learning strategy is applied to mitigate the data scarcity problem in this study. The neural network structure is TDNN-BLSTM. The multi-task TDNN-BLSTM model is illustrated as in Fig. 4. The combined layers of TDNN-BLSTM are shared among three learning tasks of phone-level acoustic modeling. The training error in a specific task is back-propagated through the corresponding task-dependent pre-final layer and the shared TDNN-BLSTM layers, while parameters for other tasks are kept unchanged. The overall cross-entropy loss function is weighted across tasks. The first task is carried out the in-domain data from unimpaired subjects in the CanAB with the goal of modeling normal speech produced on the 8 narrative tasks. Being the primary task, it is assigned the highest weight in the loss function. The second and the third learning tasks are phone-level acoustic modeling with K086 and CUSENT training data respectively. They target at modelling reading-style speech with unrestricted content.

This section describes the configurations of multi-task TDNN-BLSTM model and mono-task baseline models with the structures of feed-forward DNN, TDNN and TDNN-BLSTM. Experimental results on ASR performance evaluation are analyzed and compared in Section IX-A. Like Mandarin (Putonghua), Cantonese is a monosyllabic and tone language. Each Chinese character is spoken as a monosyllable carrying a specific tone. The Cantonese phone set contains 32 basic phone units (13 vowels and 19 consonants), 1 silence and 1 laughter, which are used as basic units for acoustic modeling.

1) Input Features: The speed-perturbation processing is applied for training data augmentation with speed factors of 0.9, 1.0 and 1.1 [33]. The Kaldi speech recognition toolkit [34] is used to extract the input features, which consist of 40-dimensional Mel-frequency cepstral coefficients (MFCCs) appended with 3-dimensional pitch features [35]. Pitch features have been shown useful in ASR of tonal languages like Cantonese [35]. The window length and window shift for short-time spectral analysis are 25 msec and 10 msec, respectively. The 43-dimensional frame-level features are spliced with a contextual window for a specific neural network model (see more details in Section V-B2 and Section V-B4). Furthermore a 100-dimensional i-vector is appended for speaker adaptation [11].

2) TDNN-BLSTM Model Structure: The TDNN-BLSTM model consists of 4 TDNN layers stacked with 4 pairs of forward-backward projected LSTM (LSTMP) layers. Each TDNN layer contains $1,024$ neurons, with ReLU activation and batch re-normalization (ReLU-renorm). LSTMP layers are with 1024-dimensional cells and 256-dimensional recurrent projections. The configurations of temporal contextual frames used at each layer are summarized in Table III, where $\{-2,-1,0,1,2\}$ indicates using 2 past frame and 2 future frame to compute an output activation. Three task-dependent pre-final layers with

TABLE III
CONTEXT CONFIGURATIONS FOR TDNN AND TDNN-BLSTM MODELS

	TDNN	TDNN-BLSTM
Layer ID	Context of TDNN layer / LSTMP layer	
1	$[-8, 8]$	$\{-2, -1, 0, 1, 2\}$
2	$\{-1, 0, 1\}$	$\{0\}$
3	$\{-1, 0, 1\}$	$\{-1, 0, 1\}$
4	$\{-3, 0, 3\}$	$\{-1, 0, 1\}$
5	$\{-6, -3, 0\}$	LSTMP-forward
6	-	LSTMP-backward
7-12	-	{LSTMP-forward, LSTMP-backward} $\times$ 3

$1,024$ neurons are implemented with ReLU-renorm activation. The output dimensions for CanAB, K086 and CUSENT are $2,496$, $2,561$ and $2,576$ respectively, which correspond to the number of distinct tri-phone states for the respective tasks.

3) TDNN-BLSTM Model Training: State-level tri-phone alignments are required as the target labels for acoustic model training. They are generated with a context-dependent GMM-HMM model trained with 40-dimensional feature-space maximum likelihood linear regression (fMLLR) features. For the training of multi-task TDNN-BLSTM, the mini-batch size is set to 64 and the number of training epochs is set to 6. The learning rate ranges from 1.5×10^{-3} to 1.5×10^{-4}, following the exponential-delay learning schedule. Dropout strategy is adopted to improve generalization ability of the acoustic model, with a probability of 0.1 [36].

4) Baseline Systems: In addition to the multi-task TDNN-BLSTM model, three single-task baseline models are implemented and evaluated. These models are DNN, TDNN and TDNN-BLSTM, which are trained with only the CanAB training set. The feedforward DNN has 6 hidden layers, each containing $1,024$ neurons with sigmoid activation functions. The input features to the DNN are sliced with a contextual window of $[-8, 8]$. The TDNN contains 5 sigmoid layers with $1,024$ neurons per layer. The context configuration of TDNN is described as in Table III. For both DNN and TDNN models, the number of training epochs is set to 3 and the learning rate is from 1.5×10^{-2} to 1.5×10^{-3} with exponential decay. For the single-task TDNN-BLSTM model, layer configurations, the mini-batch size as well as learning rate are set the same as in multi-task TDNN-BLSTM. The dropout probability is set to 0.1.

C. Language Model

To obtain a language model that matches with the aphasic speech in the CanAB, the orthographic transcription of speech utterances from control subjects in CanAB is used to train syllable tri-grams. The training is implemented with the SRILM toolkit [37].

VI. EXTRACTION OF TEXT FEATURES

Text features are designed to reflect the linguistic aspects of aphasic speech that are related to language impairments. In Law [38], narrative speech from Cantonese-speaking PWA was found to be short of amount and content, and be impoverished in structural complexity and elaboration. Kong [39] showed that aphasia speakers tend to miss more main concepts in the description of pictures than control speakers, and the fluent aphasic group performed better than the non-fluent group in producing main concepts. By carefully examining and comparing the content of selected stories from PWA and control subjects in the Cantonese AphasiaBank, the following observations are made:

- The number of topic-specific words decreases as the severity of aphasia increases;
- Sentences spoken by PWA typically are fragmented;
- Subjects and objects are often missing in aphasic speech.

Capturing topic-specific content is believed to be useful in differentiating the story told by an impaired speaker from those by unimpaired ones. The degree of discrepancies is expected to be a good indicator of the severity of aphasia. The text features are to be computed from erroneous ASR output. Two different approaches are presented in this section. Both of them are data-driven and involve the use of machine learning models.

A. Story Vector Representation

The continuous bag-of-words (CBOW) model is commonly used to map discrete words to continuous-value vector representations by capturing the relation between words [40]. In this study, a story-level vector representation is derived from word vectors to characterize the lexical and semantic content of the story. It is noted that speech utterances in the Cantonese AphasiaBank corpus contain frequent occurrences of colloquial terms and fillers, which are difficult to be transcribed into standardized Chinese characters. Therefore non-tonal syllable transcription, which is less vulnerable to ASR errors and out-of-vocabulary words, is used to represent the speech content for text feature extraction. A CBOW model is trained with syllable-level transcription of all stories from the 118 unimpaired speakers in CanAB. Here each spoken syllable of Cantonese is treated as a word, and thus the CBOW model actually learns syllable embeddings. Given a story in CanAB, the story vector is obtained by taking the average of all syllable vectors in accordance to the transcription of the story. The appropriateness and effectiveness of taking simple average are justified by the fact that basic algebraic operations can be applied to the word vectors learned by CBOW [40].

1) CBOW Training: The CBOW model is implemented using the Word2Vec Toolkit [41]. The training data is manual transcription of speech from 118 unimpaired speakers on 8 narrative tasks. The transcription contains about 183,000 syllables. The number of unique syllables is 523. A contextual window of 6 syllables is used for CBOW training. The dimension of syllable vector is set to be 50. The size of negative sampling set is set as 5 and the sub-sampling threshold is set as 10^{-3} [42].

2) Inter-Story Feature & Intra-Story Feature: We compute story vectors based on 7 tasks with specific topics (see Section III). Given story vectors from 118 unimpaired speakers and 92 impaired speakers, we extract two types of text features for each impaired speaker, namely inter-story feature and intra-story feature, following our previous work [22]. As illustrated in Fig. 5(a) (unimpaired speech) and 5(b) (impaired speech),

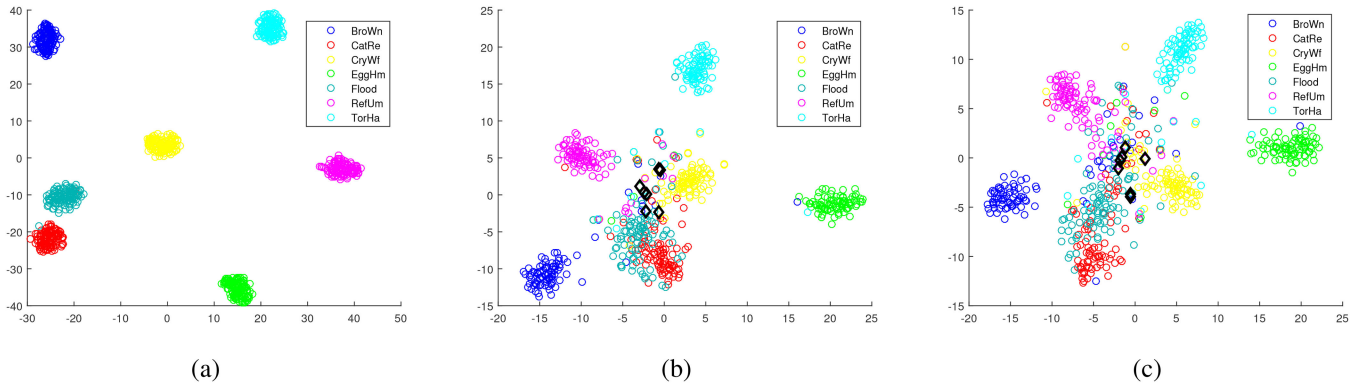


Fig. 5. 2D t-SNE visualization of story vectors based on (a) manual transcription of unimpaired speech, (b) manual transcription of impaired speech, and (c) 1-best ASR output of impaired speech. Different colors illustrate different story topics. The black diamond symbols in (b) and (c) mark the 7 story vectors from a specific speaker, who was diagnosed as Broca's aphasia with AQ value of 42.0.

the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [43] is applied to visualize the 50-dimensional story vectors derived from manual transcription in a 2-dimensional (2D) space. Different colors indicate different story topics.

For unimpaired speech, it is obviously seen that story vectors on different topics are almost perfectly separated from each other, whilst for impaired speech, a noticeable degree of overlap among story vectors can be observed. Specifically, 7 story vectors from an impaired speaker with Broca's aphasia (AQ: 42.0), marked by black diamonds "◊" in Fig. 5(b), can hardly be separated. It is due to the fact that the speech from this speaker contains mostly function words but few topic-specific content words. The inter-story and intra-story features are designed to quantify the degree of language impairment of an impaired speaker based on the significant discrepancy in the story vectors of impaired and unimpaired speakers:

- **Inter-story feature: No. of mis-clustered story vectors** The inter-story feature aims to capture the degree of confusion among 7 produced stories by counting the number of mis-clustered story vectors. Given the 7 story vectors from an impaired speaker, these vectors are first pooled with 7×118 story vectors from unimpaired speakers. K-means clustering is applied to group the pooled data into 7 classes, and the number of mis-clustered story vectors for the impaired subject is counted. The feature value is divided by 7 to be normalized in the range of 0 to 1.
- **Intra-story feature: similarity w.r.t unimpaired speech** For each of the 7 story topics, a topic vector is computed by taking the mean of the respective story vectors from all 118 unimpaired subjects. It is a compact representation of the topic as a kind of norm, against which the impaired speech would be compared. Given an impaired subject, the cosine similarity between each of the subject's story vectors and the respective topic vector is computed. The intra-story feature is defined as the average of cosine similarity measures over the 7 stories.

In practice, the story vectors have to be derived from error-prone ASR output. Fig. 5(c) illustrates story vectors of impaired speech derived from 1-best output of multi-task TDNN-BLSTM ASR system described in Section V-B2. Compared with transcription-based story vectors shown in Fig. 5(b), ASR-based story vectors are even more confused due to the ASR errors but have similar positions in the 2-dimensional space. This implies that the ASR-generated story vectors are robust to ASR errors to some extent and involve the factor of ASR performance on impaired speech for the assessment. The detailed discussion of robustness of text features to the ASR errors will be presented in Section IX-B.

B. Siamese Network: Aphasic Speech vs. Unimpaired Speech

In the previous approach, the CBOW model and the simple averaging method are utilized without considering word order. The story vectors mainly focus on reflecting the semantic content and lexical diversity of spoken stories. However, the syntactic impairment of aphasic speech, as an important indicator to the severity assessment of aphasia [6], is not taken into account. On the other hand, simply taking the mean of syllable embeddings may lead to the loss of semantic information. In this section, we propose a novel approach to extracting more comprehensive inter-story and intra-story features using siamese network.

Siamese network consists of a pair of identical sub-networks with shared weights [44]. It is able to learn high-level representations of inputs from their respective sub-networks for further comparison. The inputs for siamese network can be image data, sentences or sequential data, and thus the model is applicable to various comparison-making tasks such as image matching (e.g., face [44] and signature [45] verifications) and semantic matching (e.g., community question answering [46], off-topic response detection [47]). In this work, a siamese network is trained to compare spoken stories from impaired speakers and unimpaired ones for the text feature extraction.

1) Architecture of Siamese Network: The model architecture adopted in this study is motivated by that in [47], which is shown in Fig. 6. It aims at comparing two spoken stories A and B. Firstly, each syllable in a story is converted to a vector using the Word2Vec toolkit (see Section VI-A1). $V_1^A, V_2^A \cdots, V_L^A$ $(V_1^B, V_2^B \cdots, V_L^B)$ represent syllable vectors of story A (B) with length L. The converted vectors are concatenated to a 2-dimensional matrix with the shape of $L \times D$, where $L = 992$ denotes the maximum length among stories (padded where

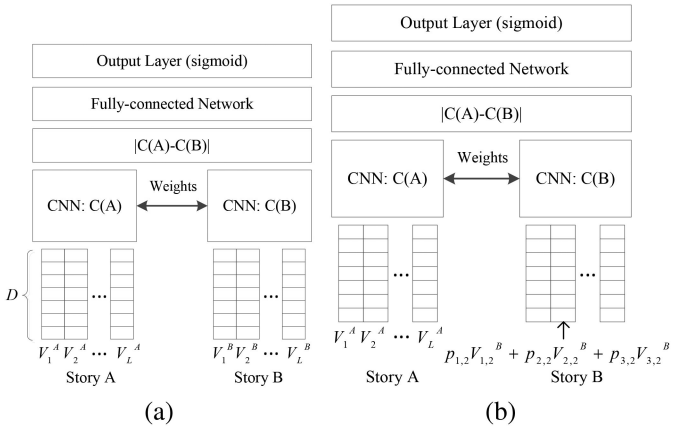


Fig. 6. Architecture of siamese network trained with (a) manual transcription or 1-best ASR output of aphasic speech, and (b) confusion networks from ASR system on aphasic speech.

necessary) and $D = 50$ is the dimension of syllable vectors. From the bottom of the siamese network, two CNNs share exactly the same architecture and weights. They are treated as descriptor computation modules to extract high-level representations from two branches of input texts. The concatenated syllable vectors are fed to 5 filter sizes of $\{3, 4, 5, 6, 7\} \times 50$ with stride 1 in the convolutional layer. Each filter size corresponds to a N-gram order, i.e., tri-grams to 7-grams. In this way, the word order is taken into consideration to reflect the syntactic impairment of aphasic speech. 100 filters are used per filter size, resulting in 500 feature maps (992×1). Max pooling is applied to capture the most important information in each feature map and the results are concatenated to a 500-dimensional vector representation. This pooling scheme naturally copes with variable sentence lengths. It is followed by a distance computation layer defined as $|C(A) - C(B)|$ (element-wise absolute difference), where $C(A)$ and $C(B)$ are vector representations of story A and story B generated from CNNs. A fully-connected layer with the size of 40 is stacked on the top of the model, followed by a sigmoid function to generate a similarity score for the pair of input stories. The ReLU activation function and dropout regularization are applied between the distance computation layer and the fully-connected layer.

2) Inter-Story Feature & Intra-Story Feature:

- **Inter-story feature: degree of confusion of 7 stories**
  Recall that the inter-story feature before is used to measure the degree of confusion among story vectors from impaired speakers using story vectors from unimpaired speakers as benchmark. Similarly, a siamese network is trained with pairs of unimpaired stories to determine whether the two spoken stories are on the same topic or not. We divide 118 unimpaired speakers into a training set and a validation set, which contain 106 speakers and 12 speakers respectively. Pairs of stories are randomly selected from 106 unimpaired speakers for training. If a pair of selected stories is on the same topic, the training target is set to 1, otherwise it is set to 0. The Area Under receiver operating characteristic Curve (AUC) is used as performance metric in the experiment (value of 1.0 indicates a perfect classification). Table IV shows that the AUCs can attain almost 1.0 for both training set and validation set, which confirms the capacity of the proposed siamese network for inter-story extraction.
  A pair of spoken stories with distinct topics from an impaired speaker are used as test data. They can be in the form of manual transcription or ASR output. There should be 21 unique pairs of stories derived from 7 stories for each impaired speaker so that 21 similarity scores can be obtained from the output of siamese network. Given an impaired subject, the mean similarity scores is defined as the inter-story feature. Compared with the inter-story feature based on story vectors (# of mis-clustered vectors), the inter-story generated from the siamese network is naturally a continuous value in the range of 0 to 1. This may be more suitable to the regression task.

TABLE IV
THE BINARY CROSS-ENTROPY LOSS AND AUC OF SIAMESE NETWORK FOR INTER-STORY FEATURE GENERATION

	Training set	Validation set
Loss	0.0011	0.0075
AUC	0.9999	1.0000

- **Intra-story feature: severity w.r.t unimpaired speech**
  For the extraction of intra-story feature, the siamese network is designed to compare the content of impaired story with that of unimpaired story within the same topic. During the training procedure, the story A and story B in Fig. 6 represent a story spoken from an unimpaired speaker and the other from an impaired speaker. They are required to be about the same topic. The training target of the siamese network is to discriminate PWA with High-AQ (AQ ≥ 90) from those with Low-AQ (AQ < 90). There are 39 subjects in the High-AQ group (label 1) and 53 subjects in the Low-AQ group (label 0). The classification label of each pair of stories is inherited from the impaired speaker. We expect that the siamese network can map the content discrepancy between impaired and unimpaired stories to the severity degree of PWA.
  The experiment on intra-story feature extraction is carried out by the five-fold cross validation strategy. As the benchmark, the text inputs of unimpaired speakers are from manual transcription. 85 unimpaired subjects are selected as training set. 24 unimpaired subjects and the rest of 9 unimpaired subjects are as test set and validation set. These three data sets are fixed in 5 folds. For input stories of impaired speakers, they can be in the form of manual transcription or ASR output. In each fold, 80% of the impaired subjects are used for training and the remaining 20% are used for test. 10% of the impaired subjects are randomly selected from training subjects as a validation set. Each of the impaired stories need to be compared with all unimpaired stories on the same topic in the respective set. The intra-story feature is defined as the average of output scores over all story pairs of an impaired speaker.

3) Hyperparameters for Model Training: The training parameters are set empirically. The mini-batch sizes are 128 for

TABLE V
COMPARISON OF PROPOSED TEXT FEATURES AND TWO BASELINE METHODS. THE CORRELATIONS ARE REPRESENTED BY THE ABSOLUTE VALUES OF SPEARMAN'S CORRELATION BETWEEN TEXT FEATURES AND AQ SCORES FOR THE 92 IMPAIRED SUBJECTS

Method	Feature	Correlation with AQ
Perplexity with N-gram	bi-grams	.517
	tri-grams	.583
	4-grams	.587
	5-grams	.580
Story vectors with bag-of-words	Inter-story	.576
	Intra-story	.728
Story vectors with syllable embeddings	Inter-story	**.647**
	Intra-story	**.829**
Siamese network with syllable embeddings	Inter-story	**.664**
	Intra-story	**.841**

training the siamese networks. The initial learning rates are set to 10^{-3}. Model training aims at minimizing the binary cross-entropy loss with the Adam optimizer [48] (weight decay coefficient 5×10^{-4}). Dropout method with probability 0.5 is used for the regularization purpose. PyTorch [49] is used for training the siamese networks in this study.

C. Effectiveness of Text Features

To examine the effectiveness of our proposed text features, we compare them with other two data-driven text features in terms of the Spearman's correlation with subjective AQ scores of 92 impaired speakers. The first one is the perplexity of N-gram model, which has been shown to be useful for automatic diagnosis of Alzheimer's disease [50]. Specifically, $\{2, 3, 4, 5\}$-gram models are also trained with syllable transcription of speech from the 118 unimpaired speakers. The perplexity is used to evaluate how well a N-gram model fits the impaired speech transcription. The other approach is based on story vectors derived from the bag-of-words model. The dimension of story vector is 523, which is equal to the number of unique syllables in the training texts (see Section VI-A1). Subsequently, the inter-story and intra-story features are computed based on these 523-dimensional story vectors.

Table V compares the absolute values of Spearman's correlation given by four approaches. All features are generated from manual transcription. It can be seen that the proposed inter-story and intra-story features perform much better than the perplexity of N-grams. N-grams only focus on the discrepancy of syllable combinations between impaired and unimpaired speech based on simple frequency count, while the inter-story and intra-story features are able to capture the content discrepancy. Another significant improvement of correlation value comes from the syllable embeddings. The syllable embedding method outperforms the bag-of-words model in generating story vectors since it can learn the semantic relation between the syllables. Fig. 7 shows the comparison between story vectors of unimpaired speakers based on bag-of-words and those based on syllable embeddings in the 2D space. It is observed that the story vectors with the same topic derived from syllable embeddings are denser than those from bag-of-words model, which confirms the advantage

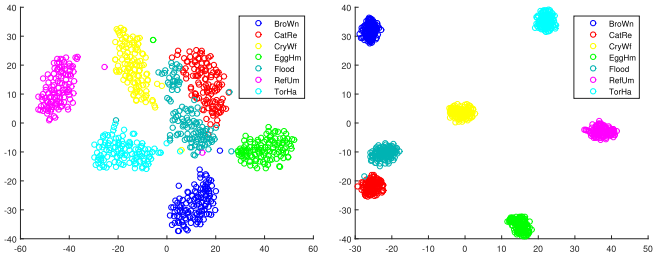


Fig. 7. 2D t-SNE visualization of manual transcription-based story vectors from unimpaired speakers. The story vectors are computed by (a) bag-of-words and (b) syllable embeddings.

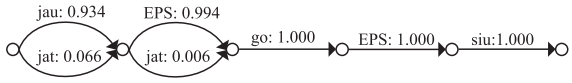


Fig. 8. Example of the confusion network of a speech segment.

of using syllable embeddings. Instead of representing a spoken story as a vector, the text features derived from a siamese network achieve the best performance. The siamese network not only learns the content of spoken stories but also has the ability to capture some syntactic information from the stories. Compared with the unsupervised story vectors, the siamese network is trained in a supervised way, which may provide additional benefits to the results.

D. Improving Feature Robustness to ASR Errors

The story vectors are desired to be derived from the error-prone ASR output in practical use. Thus, the quality of ASR output on impaired speech strongly affect the reliability of text feature extraction and further affect the accuracy of automated assessment for PWA. In the previous study [25], we incorporated rich representation of ASR output (i.e., N-best lists and confusion networks) into the computation of story vectors instead of using the straightforward 1-best ASR output. It has been demonstrated that the rich representation could provide a larger set of ASR hypotheses to facilitate more robust story vectors. In the present paper, we propose to incorporate confusion networks into text features derived from not only story vectors but also siamese network.

Confusion networks (CNs) are direct linear graphical representations of most likely hypotheses in the lattice. Fig. 8 gives an example of CNs. Each edge represents a syllable with its associated posterior probability. The summation of posterior probabilities of all candidate syllables is 1.0 at each segment. "EPS" in the CNs represents a NULL hypothesis. In the present study, we adopt the function "lattice-mbr-decode" implemented in Kaldi [34] to generate CNs.

1) Story Vectors with Confusion Networks: The story vectors incorporated with CNs are computed by the following procedures:

- Step 1. Obtain the CNs from lattices for all stories. Let L represents the length of position segments in the CN for a

TABLE VI
COMPARISON BETWEEN 13 SUPRA-SEGMENTAL DURATION FEATURES EXTRACTED BY FORCED-ALIGNMENT AND THOSE EXTRACTED BY ASR-ALIGNMENT IN TERMS OF THE SPEARMAN'S CORRELATION WITH AQ OF 92 PWA. THE FEATURES IN BOLD ARE ROBUST TO ASR ERRORS, WHICH ARE SELECTED BY A TWO-TAILED PAIRED T-TEST WITH $p > .05$

Duration Feature	Correlation with AQ		p-value
	Forced-alignment	ASR-alignment	
1. Nonspeech-to-speech duration ratio	$-.790$	$-.716$	.002
2. # of silence segments	$-.299$	$-.311$	$1.852E^{-19}$
3. Average duration of silence segments	$-.710$	$-.664$	**.420**
4. Average duration of speech segments	.777	.712	$8.160E^{-11}$
5. # of spoken syllables	.682	.638	**.284**
6. # of syllable per speech segment	.769	.727	$7.244E^{-6}$
7. Average duration per syllable	$-.310$	$-.602$	$1.588E^{-24}$
8. Ratio of silence segment count to syllable count	$-.783$	$-.740$	$3.438E^{-8}$
9. Ratio of average duration of silence to speech segments	$-.775$	$-.706$	$4.487E^{-4}$
10. Ratio of filler count to the length of speech segment	$-.507$	$-.253$	$2.529E^{-12}$
11. Syllable count per second	.733	.695	**.398**
12. # of long silence segments	$-.177$	$-.195$	$1.117E^{-20}$
13. # of short silence segments	.467	.385	$6.093E^{-6}$

story and $N_1, N_2, \cdots N_L$ denote the number of candidate syllables at L segments. For the l-th segment, the candidate syllables are $w_{1,l}, w_{2,l}, \ldots, w_{N_l,l}$ with posterior probabilities $p_{1,l}, p_{2,l}, \ldots, p_{N_l,l}$;

- Step 2. For each story, the story vector $\boldsymbol{V}_{\text{story}}$ is computed as the weighted average of all candidate syllable vectors appeared in the corresponding CN:

$$\boldsymbol{V}_{\text{story}} = \frac{\sum_{l=1}^{L}\sum_{i=1}^{N_l} p_{i,l}\boldsymbol{V}_{i,l}}{L - L_{\text{EPS}}}. \tag{1}$$

The weight for each candidate syllable corresponds to the posterior probability generated from the CN and $\boldsymbol{V}_{i,l}$ indicates the syllable vector of $w_{i,l}$. It is noted that the syllable vector for "EPS" is set as a zero vector and L_{EPS} represents the number of "EPS" with posterior probability of 1.0 in CNs. They are removed when computing the averaged story vector.

2) Siamese Network with Confusion Networks: CNs were utilized to train a BLSTM-RNN system in the spoken utterance classification task and showed better classification performance than using 1-best ASR output [51]. Motivated by the approach in this work, we propose to incorporate CNs into syllable embeddings as the input to siamese network. The modified syllable vector $\boldsymbol{V}_l^{\text{modified}}$ is given by a weighted sum representation with the posterior probabilities from CNs:

$$\boldsymbol{V}_l^{\text{modified}} = \sum_{i=1}^{N_l} p_{i,l}\boldsymbol{V}_{i,l}. \tag{2}$$

As shown in Fig. 6(b), we take the computation of V_2^{modified} as an example. The modified syllable vectors are stacked as a 2-dimensional matrix to represent a spoken story and then fed to the siamese network. The syllable vector for "EPS" is also set as a zero vector. The "EPS" with posterior probability of 1.0 in CNs is skipped when concatenating the modified syllable vectors for each input story.

VII. ACOUSTIC FEATURE EXTRACTION

Apart from linguistic impairment, symptoms like dysfluency, voice disorder and dysprosody may be present in PWA at various severity levels and with different combinations [2]. Two types of features in the acoustic aspect are extracted for developing a more comprehensive assessment system.

A. Supra-Segmental Duration Features

Supra-segmental duration features have been applied to characterize the dysfluency property of aphasic speech [22]. Based on the previous study, 13 duration features that are related to speech fluency are extracted. They are listed in Table VI. Two feature extraction approaches, relying on forced-alignment with manual transcription and time alignment with ASR decoder (ASR-alignment), are compared in terms of the Spearman's correlation with AQ values of 92 PWA.

We define the silence segments as regions of speech longer than 0.5 second and each speech segment is defined as the speech region between two silence segments. Note that parameter 1 in Table VI denotes the duration ratio between non-speech part and speech part, where the non-speech part covers filler words and silence segments while the speech part consists of all spoken syllables. The silence segments are categorized as short segments (> 0.15 second and ≤ 0.4 second) and long segments (> 0.4 second) in parameter 12 and 13 [52]. All duration features are average values computed over 9 tasks for each impaired subject.

As shown in Table VI, a number of supra-segmental duration features based on either forced-alignment or ASR-alignment show high correlations with AQ and the values reveal that the milder PWA exhibit less dysfluency. However, duration features extracted from ASR-alignment perform slightly worse than those from forced-alignment. This is reasonable due to the ASR errors on impaired speech. We find that a significant portion of the recognition errors are caused by occurrences of unintelligible speech sounds, which are not modeled by the ASR system. These sounds could be recognized as Cantonese syllables, and thus

increasing proportion of speech part and masking the degree of dysfluency.

B. eGeMAPS Features

The extended Geneva Minimslistic Acoustic Parameter Set (eGeMAPS) is a collection of acoustic features that are effective in voice-related tasks such as speech emotion recognition [53]. It has been used for automatic diagnosis of Parkinson's disease [54] and autism spectrum of children [55], which share many similarities with aphasia. The eGeMAPS contains 88 low-level acoustic features, covering spectral, cepstral, prosodic and voice quality information and can be conveniently extracted with the openSMILE toolkit [56]. These features are expected to reflect voice-related impairment in aphasic speech to facilitate the assessment of PWA.

We extract eGeMAPS features for each impaired speaker based on speech recordings of 9 tasks and compute the average value for each type of feature. For 92 impaired speakers, the absolute values of Spearman's correlations between eGeMAPS features and AQ scores range from .001 to .666.

C. Feature Selection

The goal of feature selection procedure is to select the most robust features and reduce the feature dimension for the following prediction of AQ. For the 13 candidate supra-segmental duration features, we apply a two-tailed paired t-test with $p = .05$ to select features whose values derived from ASR-alignment are not significantly different ($p > .05$) from those derived from forced-alignment. As shown in Table VI, three parameters in bold, namely "average duration of silence segments," "# of spoken syllables" and "syllable count per second," are finally selected. This suggests that the ASR performance should be further improved to obtain more accurate time alignment information.

The eGeMAPS features are directly extracted from speech recordings with no need to consider the problem of robustness. We select three of them with the highest correlations with AQ: (1) the number of loudness peaks per second (Spearman's correlation: .666); (2) the mean length of unvoiced regions (Spearman's correlation: $-.664$); (3) ratio of the energy of the spectral harmonic peak at the third formant's center frequency to the energy of the spectral peak at F0 (Spearman's correlation: .617). The selected eGeMAPS features suggest that the loudness of impaired speech, the duration of unvoiced part, and the formant information are closely related to the severity level of PWA.

VIII. Experimental Setup

The proposed system aims to automatically predict the AQ scores based on two types of text features and six types of acoustic features including three supra-segmental duration features and three eGeMAPS features. All 8 features are z-normalized separately in order to make their values in the comparable range. They are combined as an 8-dimensional feature vector for each impaired subject. Automatic prediction of AQ is formulated by a regression task. Two different regression models based on linear regression (LR) and random forest (RF) are applied and evaluated. Leave-one-out cross-validation strategy is adopted for performance evaluation. In each fold, feature vectors from 91 impaired speakers are used to train the regression models and the feature vector from the remaining one impaired speaker is used as the test data. As a result, a predicted AQ score is obtained for each of the 92 aphasia speakers.

TABLE VII
Different Schemes of Feature Combination for Performance Evaluation

Scheme	Text Features	Duration Features
Trans-storyvec	transcription+story vectors	forced-alignment
Trans-siamese	transcription+siamese network	forced-alignment
ASR-1best-storyvec	1-best ASR output+story vectors	ASR-alignment
ASR-CN-storyvec	CNs+story vectors	ASR-alignment
ASR-1best-siamese	1-best ASR output+siamese network	ASR-alignment
ASR-CN-siamese	CNs+siamese network	ASR-alignment

The experiments are carried out using six different schemes of feature combination as shown in Table VII. The eGeMAPS features extracted from raw speech recordings are identical across the 6 schemes. The regression models are trained independently for each scheme. The experiments are designed to compare the performance of AQ prediction in the ideal case of using a perfect ASR (manual transcription) with that in the practical case of using an imperfect ASR.

IX. Results and Discussion

A. ASR Performance

The performance of baseline systems and multi-task TDNN-BLSTM system are compared in Table VIII. They are evaluated on both *Aphasia* group (92) and *Control* group (17) in the test set of CanAB. We use syllable error rate (SER) as the performance metric. The overall SER is denoted as "SER" in Table VIII.

For acoustic models of DNN, TDNN and TDNN-BLSTM trained with CanAB using mono-task learning, TDNN-BLSTM model outperforms other models. Using TDNN-BLSTM acoustic model significantly reduces overall SER by 5.79% and 3.40% compared with TDNN model for the *Aphasia* group and the *Control* group respectively. This implies that the long contextual information captured by BLSTM structure is important to aphasic speech recognition. The multi-task TDNN-BLSTM achieves the best performance among all acoustic models, with relative improvements of SER of 5.66% for the *Aphasia* group and 3.40% for the *Control* group compared with its counterparts in mono-task case. The result shows that a large amount of speech data with different speaking styles can be jointly learned to improve the generalization capability of the acoustic model under the multi-task framework.

The speaker-level SER and story-level SER of *Aphasia* group and *Control* group are further compared. Table IX summaries the mean and standard deviation of speaker-level SER and story-level SER for two groups of subjects. Compared with the *Control* group, the SER per speaker varies greatly for the *Aphasia* group, reflecting the highly diverse types and degrees

TABLE VIII
PERFORMANCE COMPARISON BETWEEN BASELINE AND MULTI-TASK TDNN-BLSTM SYSTEMS (WITH OPTIMIZED TASK WEIGHTS) ON CANAB TEST SET. OVERALL SER (SER) AND SER ON SPEECH OF 7 TASKS (SER-7) ARE USED AS PERFORMANCE METRICS

Acoustic Model	Training Dataset(s): weights	Test Data	SER (%)	SER-7 (%)
DNN	CanAB	Aphasia	46.56	44.51
		Control	19.46	16.85
TDNN	CanAB	Aphasia	46.46	44.82
		Control	19.00	16.50
TDNN-BLSTM	CanAB	Aphasia	40.67	39.07
		Control	15.60	13.47
Multi-task TDNN-BLSTM	CanAB, K086, CUSENT: 0.65, 0.2, 0.15	Aphasia	**38.37**	**37.12**
		Control	**15.07**	**13.16**

TABLE IX
MEAN AND STANDARD DEVIATION OF SPEAKER-LEVEL SER AND STORY-LEVEL SER FOR *APHASIA* GROUP (92) AND *CONTROL* GROUP (17), WHERE STORY-LEVEL SER IS COMPUTED OVER 9 TASKS FOR *APHASIA* GROUP AND 8 TASKS FOR *CONTROL* GROUP

Type	Aphasia	Control
Speaker-level SER (%)	46.45 ± 20.73	15.18 ± 6.91
Story-level SER (%)	39.34 ± 2.71	14.99 ± 3.90

TABLE X
DEVIATION OF TEXT FEATURES COMPUTED FROM ASR OUTPUT FROM THOSE FROM MANUAL TRANSCRIPTION. THE IMPAIRED SPEAKERS ARE DIVIDED INTO TWO GROUPS ACCORDING TO SER. THE AVERAGE DEVIATION ACROSS ALL SPEAKERS IN EACH GROUP IS SHOWN. THE SMALLEST ABSOLUTE DEVIATION VALUE FOR INTER-STORY FEATURE AND INTRA-STORY FEATURE IN EACH GROUP IS MARKED IN RED AND BLUE RESPECTIVELY

Method	Text feature	Deviation of feature values	
		SER $\leq 50\%$ (59)	SER $> 50\%$ (33)
ASR-1best-storyvec	Inter-story	0.012	0.091
	Intra-story	0.005	−0.060
ASR-CN-storyvec	Inter-story	**0.002**	0.074
	Intra-story	0.009	**−0.047**
ASR-1best-siamese	Inter-story	0.003	**0.038**
	Intra-story	0.025	−0.052
ASR-CN-siamese	Inter-story	**0.002**	0.041
	Intra-story	**0.001**	−0.105

of language impairment. Across different tasks, the ASR system shows similar variation in performance for both groups. For the impaired speakers, two tasks of personal monologue named "Stroke" and "ImpEv" obtain top two SERs with the percentages of 43.75% and 42.82% among 9 tasks. The spoken content and vocabulary of personal monologue are not within a specific topic, and thus the language model finds it hard to deal with the unseen domains. Table VIII also shows the SER of speech from 7 tasks (except two personal monologues), which is denoted as "SER-7". As expected, a SER decrease can be seen for both groups. The ASR output of impaired speech of 7 tasks (SER-7: 37.12%) is used to extract text features for aphasic speech assessment.

For impaired speakers, the correlation between their overall SERs and subjective AQ scores is of great interest to us. We expect that the speech from more severe PWA is more likely to be mis-recognized due to impaired fluency and frequent non-speech events. The Spearman's correlation of $-.603$ confirms a relatively high correlation between SERs and AQ scores, suggesting that the ASR performance is probable to reflect the severity degree of PWA. However, the SER information cannot be obtained without manual transcription so that it is limited to some specific applications.

B. Robustness of Text Features to ASR Errors

The robustness of text features to ASR errors is critical to realizing the ASR-driven assessment system. In this section, we will examine the effectiveness of CNs in improving feature robustness and investigate the effect of ASR errors on the assessment results.

As shown in Table X, we divide the 92 impaired speakers into two groups based on the SER: below 50% vs. over 50%. There are 59 impaired subjects with SER lower than 50% while 33 speakers are with SER higher than 50%. For each group, we compare the average deviation of text features computed with the ASR output from those with manual transcription ($text\ feature_{\text{ASR}} - text\ feature_{\text{manual}}$). Smaller absolute value of the distance indicates that the ASR-generated text features is closer to those based on manual transcription. For the inter-story feature, a positive deviation means that the degree of confusion of stories (e.g., # of mis-clustered story vectors, content similarity) tend to be over-estimated based on the ASR output, hence the degree of impairment would be over-estimated. For the intra-story feature (e.g., cosine similarity, severity score), a negative deviation indicates over-estimated discrepancy between impaired and unimpaired content, and would also lead to the over-estimation of impairment severity.

For the group of low-SER subjects, the text features computed with CNs, namely the ASR-CN-storyvec and ASR-CN-siamese, are more robust to ASR errors than those computed with 1-best ASR output in most cases. The smallest average deviation of inter-story feature is 0.002 in ASR-CN-storyvec and ASR-CN-siamese, which can be treated as an over-count of 0.014 mis-clustered story vectors (out of 7). The average deviation of intra-story feature attained by ASR-CN-siamese is also very small (0.001). For the high-SER subjects, the intra-story and inter-story text features derived from ASR-CN-storyvec and ASR-1best-siamese achieve the best performance. Compared with the subjects with low SER, the text features computed from ASR output deviate more noticeably. This reveals that poor ASR performance accounts for the over-estimation of impairment

TABLE XI
AQ PREDICTION PERFORMANCE UNDER SIX EXPERIMENTAL SCHEMES IN TERMS OF THE SPEARMAN'S CORRELATION

Method	Correlation with AQ	
	Linear Regression	Random Forest
Trans-storyvec	.833	.837
Trans-siamese	**.843**	**.844**
ASR-1best-storyvec	.804	.808
ASR-CN-storyvec	.806	.815
ASR-1best-siamese	**.823**	.825
ASR-CN-siamese	**.823**	**.827**

severity. With the approach of story vectors, using the CNs can improve the robustness of inter-story and intra-story features in terms of the deviation values for high-SER group. Whilst with the siamese network, using the 1-best ASR output achieves smaller absolute deviation values than using the CNs.

Overall speaking, the ASR errors would cause over-estimation of severity level and the robustness of text features can be improved with the incorporation of CNs, especially for subjects with low SER.

C. Automatic Prediction of AQ

With the LR and RF models, the prediction performance of six experimental schemes are compared in Table XI. Transcription type (manual transcription vs. ASR output), approach to text feature extraction (story vectors vs. siamese network), approach to feature robustness improvement (1-best ASR output vs. CNs) and regression model (LR vs. RF) are considered in the comparison. Table XI summarizes the AQ prediction results which are measured in Spearman's correlation between the predicted AQ (AQ_p) and the reference AQ (AQ_r) scores of 92 PWA. Not surprisingly, features derived from manual transcription would lead to more accurate prediction results than those from ASR output, with the best correlation value of .844 vs. .827. This suggests that the automatic assessment system still have the potential to be further improved. The siamese network significantly outperforms the story vectors in generating text features for AQ score prediction, whether the text features are derived from manual transcription or ASR output, reflecting that the siamese network is more suitable than the story vectors in this specific assessment task. With the help of CNs for improving the robustness of text features, the AQ prediction performance can be accordingly enhanced, but the improvement is not statistically significant. In addition, it can be seen that the RF model performs better than the LR model in the AQ prediction task.

Table XII summarizes AQ prediction performance based on individual feature groups, namely 2-dimensional text features, 3-dimensional duration features and 3-dimensional eGeMAPS features, using the RF regression model. Manual transcription-based, ASR-based text features and duration features are compared. The results show that manual transcription-based features are more effective than ASR-based ones. The text features are found to be more useful in AQ prediction than the duration and eGeMAPS features, suggesting that language impairment is more significant than acoustic impairment in aphasic speech.

TABLE XII
AQ PREDICTION PERFORMANCE BASED ON INDIVIDUAL FEATURE GROUPS USING RF REGRESSION MODEL, MEASURED IN SPEARMAN'S CORRELATION

Feature type (dimension)	Derived from	Correlation with AQ
Text features-storyvec (2)	Transcription	.817
	ASR-1best	.814
	ASR-CN	.814
Text features-siamese (2)	Transcription	.844
	ASR-1best	.815
	ASR-CN	.825
Duration features (3)	Froced-alignment	.741
	ASR-alignment	.711
eGeMAPS features (3)	Raw speech	.665

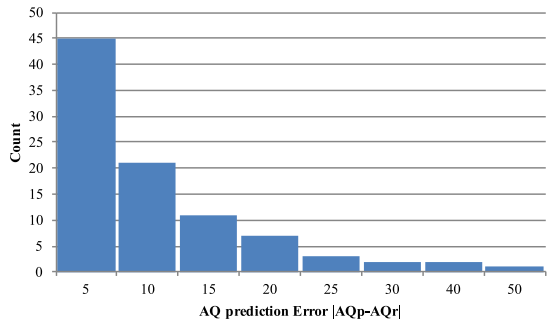


Fig. 9. Histogram of AQ prediction errors $|AQ_p - AQ_r|$.

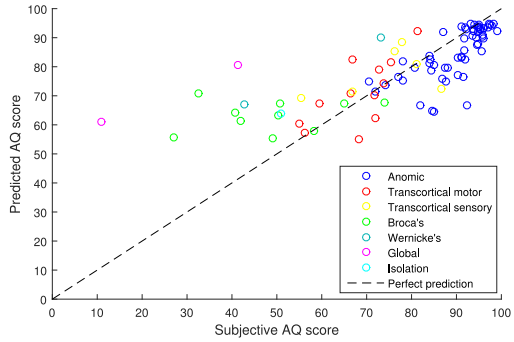


Fig. 10. The scatter plot of predicted AQ versus subjective AQ. Each circle symbol corresponds to one aphasia speaker. Different aphasia types are represented by different colors.

Compared with the results in Table XI, the combination of all proposed features shows better performance than individual feature groups in most cases. Therefore, considering multiple aspects of aphasic speech is critical to the design of an assessment system.

The best automatic assessment system attains a correlation of .827 using the RF regression model under the scheme of ASR-CN-siamese. With this model, we plot the histogram of AQ prediction errors $|AQ_p - AQ_r|$ in Fig. 9. Almost a half, i.e., 48.9% $(45/92)$ of impaired speakers obtain the prediction errors $|AQ_p - AQ_r| \leq 5.0$, and 71.7% $(66/92)$ have the prediction errors smaller than 10.0. The result demonstrates that the 8 proposed features are fairly reliable in predicting the subject AQ scores. Fig. 10 gives a scatter plot of predicted AQ versus

ground-truth AQ. It is found that AQ prediction is more accurate for aphasia speakers with high AQ scores. These speakers represent mainly the types of Anomic, Transcortical motor and Transcortical sensory aphasia. For subjects with AQ < 50, the predicted AQs tend to be higher than subjective AQs. This may be related to imbalanced data distribution that only 8 speakers are with AQ < 50. It is expected that the quality of ASR output would affect the assessment accuracy. Therefore, we compute the correlation between SERs and AQ prediction errors of 92 PWA but the correlation value of .339 is not significant. A similar result was found in an automatic assessment system for English-speaking PWA [11].

To investigate the possible causes of prediction errors, we analyze two typical impaired subjects whose AQ prediction errors are greater than 10.0. The SER of these two selected speakers are on the low side, i.e., 27.89% and 24.60%, ensuring that the text features are not much affected by ASR errors. The AQ value of the former subject is underestimated ($AQ_p = 77.1$ vs. $AQ_r = 90.4$). We find that there are frequent onomatopoeia words in his speech, such that the value of computed intra-story text feature is significantly small. It should be noted that the AQ is a composite score measuring multiple aspects of speech impairment of PWA. Some parts of the sub-tests, e.g., auditory verbal comprehension, naming, are not related to the speech impairments that the proposed features aim to characterize. It is possible that the subject performed very well in most of the sub-tests and obtained a high combined score, but was not able to handle the narrative tasks [57].

The other selected impaired subject has an over-estimated predicted AQ score, i.e., $AQ_p = 90.2$ vs. $AQ_r = 73.2$. The text features of this subject are relatively good, meaning that the content of her spoken stories is quite relevant to the given topics. However, the syntactic constructions of her spoken sentences are often incomplete and confused. In addition, topic-specific keywords and long phrases were repeated several times when she tried to produce the next sentence. This over estimation problem reveals that our proposed model is not able to sufficiently capture the syntactic characteristics of aphasic speech. If the subject's speech includes a great number of topic-specific words, the resulted text features would not accurately reflect the severity level of PWA. We consider it necessary to further explore other text features to characterize syntactic impairment in aphasic speech in the following study.

X. Conclusion

In this paper, a fully automated assessment system based on narrative speech from Cantonese-speaking PWA is presented. It has been demonstrated that the proposed data-driven text features are very effective in detecting language impairment in aphasic speech. Text features learned by the siamese network show the highest correlation with subjective AQ scores. By leveraging confusion network as enriched representation of ASR output, the robustness of text features could be further improved. Combined with other acoustic features, a high correlation of .827 can be achieved between predicted AQ and reference AQ. The proposed system provides a more efficient way of processing and analyzing pathologically impaired speech for the purposes of diagnosis and rehabilitation.

For follow-up work, there is clearly a need to improve the performance of ASR on aphasic speech so as to produce more robust features. It is also important to apply our proposed approach to other databases of pathological speech and different languages. Finally, automatic classification of aphasia type is highly desirable from the clinical perspective, and this requires a substantial effort of large-scale data collection.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. They would like to acknowledge the contribution of the people living with aphasia who participated. They are also grateful to all research assistants and student helpers for their work on the annotation and pre-processing of the speech data in Cantonese AphasiaBank.

References

[1] D. F. Benson, D. F. Benson, and A. Ardila, *Aphasia: A Clinical Perspective*. Oxford, U.K.: Oxford Univ. Press, 1996, pp. 89–98.

[2] Wikipedia Contributors, "Aphasia—Wikipedia, the free encyclopedia," 2018. [Online]. Available: https://en.wikipedia.org/wiki/Aphasia. Accessed on: Sep. 10, 2018.

[3] American Speech-Language-Hearing Association, "Signs and symptoms," 2019. [Online]. Available: https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589934663&section= Signs_and_Symptoms. Accessed on: Feb. 16, 2019.

[4] N. Simmons-Mackie, A. Raymer, E. Armstrong, A. Holland, and L. R. Cherney, "Communication partner training in aphasia: A systematic review," *Arch. Phys. Med. Rehabil.*, vol. 91, no. 12, pp. 1814–1837, Dec. 2010.

[5] A. Kertesz, *WAB-R: Western Aphasia Battery-Revised*. San Antonio, TX, USA: PsychCorp, 2007.

[6] E. M. Yiu, "Linguistic assessment of Chinese-speaking aphasics: Development of a Cantonese Aphasia Battery," *J. Neurolinguist.*, vol. 7, no. 4, pp. 379–424, Oct. 1992.

[7] R. Prins and R. Bastiaanse, "Analyzing the spontaneous speech of aphasic speakers," *Aphasiology*, vol. 18, no. 12, pp. 1075–1091, Jan. 2004.

[8] K. C. Fraser, F. Rudzicz, and E. Rochon, "Using text and acoustic features to diagnose progressive aphasia and its subtypes," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2177–2181.

[9] K. C. Fraser *et al.*, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, vol. 55, pp. 43–60, Jun. 2014.

[10] K. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proc. Workshop Speech Lang. Process. Assist. Technol.*, Grenoble, France, 2013, pp. 47–54.

[11] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Commun.*, vol. 100, pp. 1–12, Jun. 2018.

[12] H. Stadthagen-Gonzalez and C. J. Davis, "The Bristol norms for age of acquisition, imageability, and familiarity," *Behav. Res. Methods*, vol. 38, no. 4, pp. 598–605, Nov. 2006.

[13] K. J. Gilhooly and R. H. Logie, "Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words," *Behav. Res. Methods Instrum.*, vol. 12, no. 4, pp. 395–427, Jul. 1980.

[14] T. Lee *et al.*, "Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, 2016, pp. 6475–6479.

[15] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. L. G. Tempini, and J. Ogar, "Learning diagnostic models using speech and language measures," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, Vancouver, BC, Canada, 2008, pp. 4648–4651.

[16] C. Kohlschein and D. Klischies, "Automatic processing of clinical aphasia data collected during diagnosis sessions: Challenges and prospects," in *Proc. Workshop RaPID-2 Int. Conf. Lang. Resour. Eval.*, Miyazaki, Japan, 2018, pp. 11–18.

[17] Y. Qin, Y. Wu, T. Lee, and A. P. H. Kong, "An end-to-end approach to automatic speech assessment for Cantonese-speaking people with aphasia," Mar. 2019, *arXiv:1904.00361*.

[18] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Comput. Speech Lang.*, vol. 27, no. 6, pp. 1147–1162, Sep. 2013.

[19] H. Christensen *et al.*, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Proc. Interspeech*, Lyon, France, 2013, pp. 3642–3645.

[20] H. Christensen, P. D. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech," in *Proc. Interspeech*, Lyon, France, 2013, pp. 1159–1163.

[21] Y. Liu, Y. Qin, S. Feng, T. Lee, and P. Ching, "Disordered speech assessment using Kullback-Leibler divergence features with multi-task acoustic modeling," in *Proc. IEEE Int. Symp. Chin. Spoken Lang. Process.*, Taipei, Taiwan, 2018, pp. 61–65.

[22] Y. Qin, T. Lee, and A. P. H. Kong, "Automatic speech assessment for aphasic patients based on syllable-level embedding and supra-segmental duration features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, 2018, pp. 5994–5998.

[23] D. Le, K. Licata, and E. M. Provost, "Automatic paraphasia detection from aphasic speech: A preliminary study," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 294–298.

[24] D. Le and E. M. Provost, "Improving automatic recognition of aphasic speech with AphasiaBank," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 2681–2685.

[25] Y. Qin, T. Lee, S. Feng, and A. P. H. Kong, "Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 3418–3422.

[26] A. P.-H. Kong and S.-P. Law, "Cantonese AphasiaBank: An annotated database of spoken discourse and co-verbal gestures by healthy and language-impaired native Cantonese speakers," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1131–1144, Jun. 2019.

[27] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, Sep. 2011.

[28] A. P.-H. Kong, S.-P. Law, C. C.-Y. Kwan, C. Lai, and V. Lam, "A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (DoSaGE)," *J. Nonverbal Behav.*, vol. 39, no. 1, pp. 93–111, Mar. 2015.

[29] B. MacWhinney, *The CHILDES Project: The Database*, vol. 2. Hove, U.K.: Psychology Press, 2000.

[30] P. Ching, T. Lee, W. Lo, and H. Meng, "Cantonese speech recognition and synthesis," in *Advances in Chinese Spoken Language Processing*. Singapore: World Scientific, Dec. 2006, pp. 365–386.

[31] SpeechOcean, "King-ASR-086," 2011. [Online]. Available: http://en.speechocean.com/datacenter/details/146.html

[32] T. Lee, W. K. Lo, P. Ching, and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Commun.*, vol. 36, no. 3/4, pp. 327–342, Mar. 2002.

[33] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3586–3589.

[34] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Big Island, HI, USA, 2011.

[35] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 2494–2498.

[36] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with LSTMs," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 1586–1590.

[37] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, USA, 2002, pp. 901–904.

[38] S. Law, "A quantitative analysis of Cantonese aphasic production," *J. Psychol. Chin. Soc.*, vol. 2, no. 2, pp. 211–237, 2001.

[39] A. P.-H. Kong, "The use of main concept analysis to measure discourse production in Cantonese-speaking persons with aphasia: A preliminary report," *J. Commun. Disorders*, vol. 42, no. 6, pp. 442–464, Jun. 2009.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, Scottsdale, AZ, USA, 2013.

[41] Google Inc., "Word2vec," 2013. [Online]. Available: https://code.google.com/archive/p/word2vec/

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2013, pp. 3111–3119.

[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov., pp. 2579–2605, Nov. 2008.

[44] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 539–546.

[45] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 1994, pp. 737–744.

[46] A. Das, H. Yenala, M. Chinnakotla, and M. Shrivastava, "Together we stand: Siamese networks for similar question retrieval," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, Berlin, Germany, 2016, vol. 1, pp. 378–387.

[47] C. M. Lee, S.-Y. Yoon, X. Wang, M. Mulholland, I. Choi, and K. Evanini, "Off-topic spoken response detection using Siamese convolutional neural networks," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 1427–1431.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, 2015.

[49] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NeurIPS Autodiff Workshop*, Long Beach, CA, USA, 2017.

[50] S. Wankerl, E. Nöth, and S. Evert, "An N-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3162–3166.

[51] R. Masumura, Y. Ijima, T. Asami, H. Masataki, and R. Higashinaka, "Neural ConfNet classification: Fully neural network based spoken utterance classification using word confusion networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, 2018, pp. 6039–6043.

[52] S. V. Pakhomov *et al.*, "Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration," *Cogn. Behav. Neurol.*, vol. 23, no. 3, pp. 165–177, Sep. 2010.

[53] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

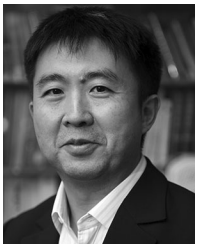
[54] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson's disease diagnosis using machine learning and voice," in *Proc. IEEE Signal Process. Med. Biol. Symp.*, Philadelphia, PA, USA, 2018, pp. 1–7.

[55] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, "Speech-based diagnosis of autism spectrum condition by generative adversarial network representations," in *Proc. Int. Conf. Digit. Health*, London, U.K., 2017, pp. 53–57.

[56] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[57] A. P. H. Kong, *Analysis of Neurogenic Disordered Discourse Production: From Theory to Practice*. New York, NY, USA: Routledge, 2016.

Ying Qin (S'16) received the B.Eng. degree in electronic and information engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2015. She is currently working toward the Ph.D. degree in electronic engineering with the DSP and Speech Technology Laboratory, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong. Her research interests are mainly on automatic speech assessment, automatic speech recognition, and machine learning with applications on pathological speech.



Tan Lee is currently an Associate Professor and the Director of Undergraduate Studies with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong. He has been working on speech and language related research for many years. His research covers spoken language technologies, speech enhancement and separation, audio and music processing, speech and language rehabilitation, and neurological basis of speech and language. He led the effort on developing Cantonese-focused spoken language technologies that have been widely licensed for industrial applications. His current work is focused on applying signal processing and machine learning methods to atypical speech and language that are related to different kinds of human communication and cognitive disorders. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the *EURASIP Journal on Advances in Signal Processing*. He is the Vice Chair of ISCA Special Interest Group of Chinese Spoken Language Processing, and an Area Chair in the technical program committees of Interspeech 2014, 2016, and 2018.

Anthony Pak Hin Kong is currently an Associate Professor with the School of Communication Sciences and Disorders, University of Central Florida, Orlando, FL, USA. His research interests include Chinese aphasia, discourse analyses, and neurogenic communication disorders in multilingual speakers. Over the years, he has developed a range of clinical assessment batteries of language and cognition geared toward Chinese speakers, including the Cantonese version of Birmingham Cognitive Screen, The Main Concept Analysis for oral discourse production, and The Hong Kong version of the Oxford Cognitive Screen. He has served as a Consultant to provide research, clinical, and/or professional consultations to international agencies, such as Aphasia United, Hong Kong Hospital Authority, Self Help Group for the Brain Damaged, Hong Kong Association of Speech Therapists, Hong Kong Society for Rehabilitation, and Hong Kong Productivity Council. He is also an Academic Editor of *PLOS ONE*.