Automatic Speech Assessment for Aphasic Patients Based on Syllable-Level Embedding and Supra-Segmental Duration Features

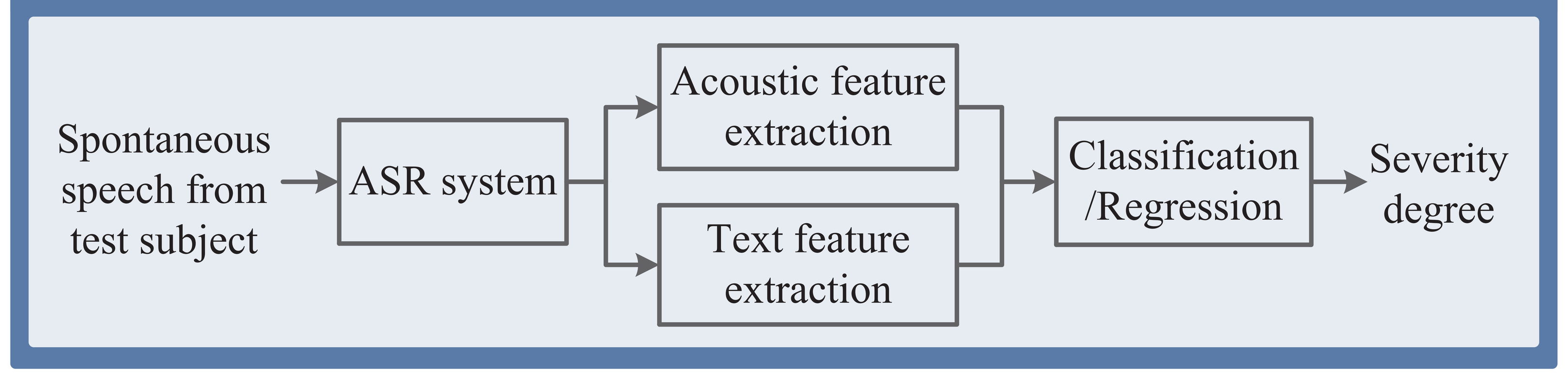
Ying Qin[1], **Tan Lee**[1] and **Anthony Pak Hin Kong**[2]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong [2]Department of Communication Sciences and Disorders, University of Central Florida

Background & Motivation

- **Aphasia**: acquired language impairment caused by brain injury.
  – Affecting phonology, lexicon, syntax, semantics of language system.
- **Speech assessment**: essential part of aphasia assessment.
  – Determine severity/type of impairment.
  – **Acoustical** and **linguistic** analysis of **story-telling speech**.
- **Subjective assessment**: by speech pathologists.
  – Requiring clinical, linguistic and cultural background knowledge.
- **Goal**: automatic speech assessment for people with aphasia (PWA).

Framework for Fully Automatic Assessment



- **Contributions**:
  – **Robust text features** with **word embedding techniques**.
  – ASR-derived features. → **manual transcription not needed**.

Dataset: Cantonese AphasiaBank

- **Spontaneous speech**: 104 aphasic and 149 unimpaired subjects.
- **Narrative tasks**: 4 picture descriptions, 1 procedure description, 2 story telling and 1 personal monologue.
- **"Story"**: except the personal monologue, the speech of each task is about a specific topic. → 7 stories.
- **Subjective score**: based on Cantonese Aphasia Battery.
  – Aphasia Quotient (**AQ**: 0-100). Low → severe.

ASR System for Aphasia Assessment

- **Target**: domain- and speaking-style-matched ASR system.
- **Training set**: 12.6h speech recordings of 101 unimpaired speakers.
- **Test set**: speech recordings of 17 unimpaired speakers and 82 aphasic speakers on 7 stories.
- **Acoustic model**:
  – Structure: standard DNN-HMM with 6 hidden layer and 1024 neurons per layer.
  – Pronunciation lexicon: 630 Cantonese syllables.

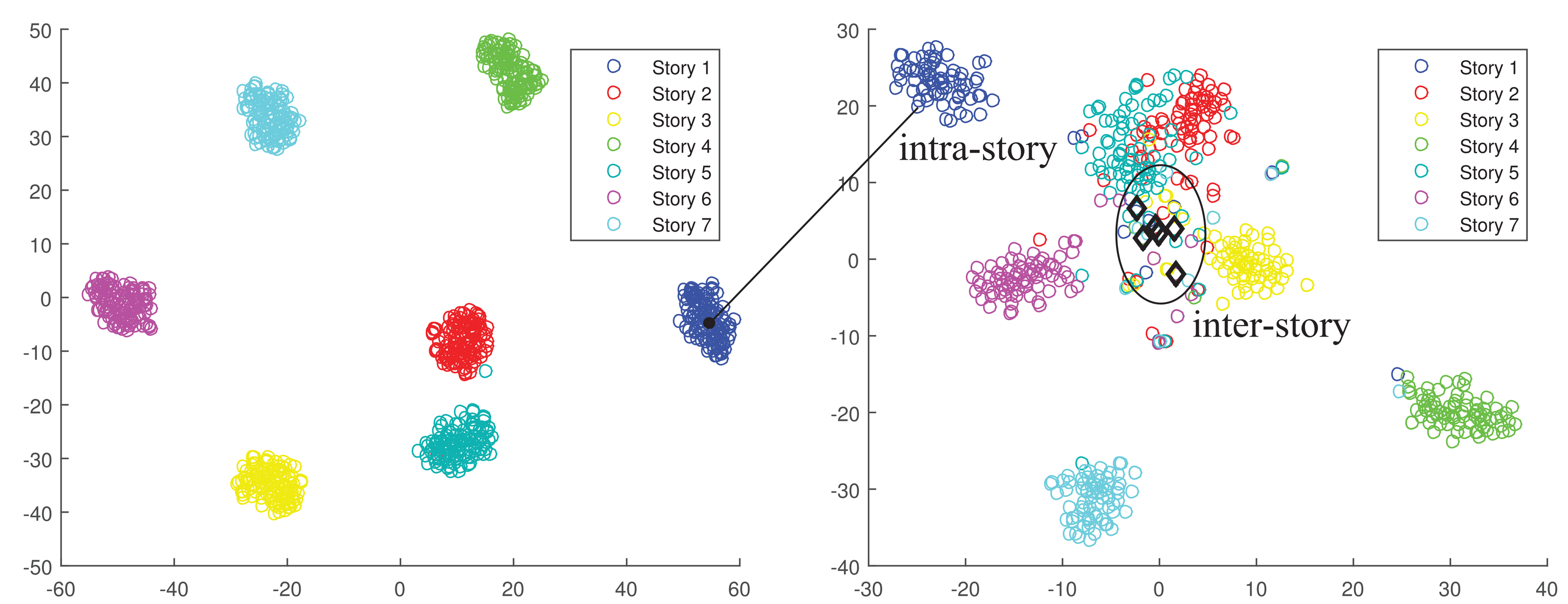
Syllable = **Initial** + **Final**
e.g. **tek bo** 踢球

  – Acoustic unit: 20 Initials + 53 Finals + 5 non-content sounds.
- **Language model**: syllable bi-grams.
- **Implementation**: Kaldi Speech Recognition Toolkit.
- **Performance**: syllable error rate (SER).
  – Unimpaired speakers: 18.24% vs. impaired speakers: 48.08%.

Feature Extraction

1. Text Features: Syllable-level Embedding Features

- **Observation**: No. of topic-specific keywords decrease; subjects and objects are missing in impaired sentences.
- **Goal**: robust text features that reflect topic-specific content of a story and differentiate unimpaired story and impaired ones.
- **Method**: 50-dimensional story-level vector representation.
  – CBOW model trained by syllable-level transcriptions of unimpaired speech.
  – Story vector: average of syllable vectors in the story.
- **Implementation**: Word2vec Toolkit.
- 2D display of story vectors derived from manual transcriptions:



(a) unimpaired speech (b) impaired speech

- **Two types of text features:**
- **Inter-story feature**: No. of misclustered story vectors (divide 7).
  – K-means clustering: 7 story vectors from an impaired subject + 7 × 118 story vectors from unimpaired subjects. → 7 classes.
  – Degree of content confusion. → Few content words.
- **Intra-story feature**: cosine similarity between an impaired story vector and the unimpaired story vectors (mean) on the same topic.
  – Discrepancy between impaired and unimpaired content.
- **Effect of ASR performance**: average deviation (text feature$_{ASR}$−text feature$_{transcription}$) for low- and high-SER groups.

SER		SER ≤ 50%	SER > 50%
Deviation of feature values	Inter-story	0.020	0.182
	Intra-story	0.002	-0.092

– Deviation is small for low-SER group.
– Deviation is more noticeable for high-SER group.
– Over-estimation of impairment severity for high-SER group.

2. Acoustic Features: Supra-segmental Duration

- **Observation**: impaired fluency of speech.
- **Method**: 13 features from syllable-level time alignment of ASR.
- **Feature selection**: LASSO regression + correlation with AQ.
- **Five types of acoustic features:**
  - **Non-speech-to-speech duration ratio**.
  - **Average duration of silence segments** (>0.5s).
  - **Average duration of speech segments**.
  - **Ratio of silence segment count to syllable count**.
  - **Syllable count per second**.

Experimental Results

1. Binary Classification of Aphasia Severity

- High-AQ: AQ>90 (35) vs. Low-AQ: AQ<90 (47).
- Leave-one-out cross validation.
- Binary decision tree (BDT), random forest (RF), and support vector machine (SVM).
- Classification results in F1 score:

	BDT	RF	SVM
Text features only (2)	0.851	0.896	0.841
Acoustic features only (5)	0.792	0.821	0.789
All features (7)	**0.891**	**0.903**	**0.874**

- With the best classifier, the recalls for Low-AQ and High-AQ groups are 89.4%(42/47) and 88.6%(31/35).

2. Automatic Prediction of AQ

- Regression problem.
- Leave-one-out cross validation.
- Linear regression (LR) and random forest (RF).
- Correlations between predicted AQ (AQ_p) and reference AQ (AQ_r):

	LR	RF
Text features only (2)	0.821	0.820
Acoustic features only (5)	0.651	0.655
All features (7)	0.816	**0.839**

- 50% (41/82) with $|AQ_p$-$AQ_r|<5.0$, 74.4% (61/82) smaller than 10.0.
- **Summary**:
  – Text features are more effective.
  – Two types of features are complementary to each other.

Future work

- Improve the ASR accuracy on impaired speech.
- Explore other text features related to syntactic impairment.
- Analyze relation between aspects of AQ score and proposed features.
- Enlarge the PWA speech database.