COMBINING PHONE POSTERIORGRAMS FROM STRONG AND WEAK RECOGNIZERS FOR AUTOMATIC SPEECH ASSESSMENT OF PEOPLE WITH APHASIA

Ying Qin[†]*, Tan Lee*[†] *and Anthony Pak Hin Kong*[‡]

[†] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[‡] Department of Communication Sciences and Disorders, University of Central Florida, Orlando, FL, USA
[†]yingqin@link.cuhk.edu.hk, [†]tanlee@cuhk.edu.hk, [‡]antkong@ucf.edu

ABSTRACT

This paper presents an investigation on applying automatic speech recognition (ASR) to speech assessment of people with aphasia (PWA). A distinctive characteristic of PWA speech is paraphasia, which refers to frequent occurrence of phonemic errors, unintended words and non-verbal sounds. In view of the wide variety of paraphasias, we propose to view the ASR errors so caused as out-of-vocabulary (OOV) words. Inspired by previous research on OOV detection, paraphasias in PWA speech are captured by comparing the phone posteriorgrams of a strongly constrained speech recognizer and a weakly constrained one. The posteriorgrams also reveal other characteristics of impaired speech, e.g., change of speaking rate, voice abnormality. Siamese and 2-channel convolutional neural network (CNN) models are used for classifying the posteriorgram pairs and predicting the severity of aphasia. Experimental results on a Cantonese database of PWA speech confirm the effectiveness of the proposed methods. The best F1 score attained on binary classification (severe versus mild aphasia) is 0.891.

***Index Terms*—** Aphasia, speech assessment, ASR, phone posteriorgrams, CNN

1. INTRODUCTION

Aphasia is an acquired neurogenic speech-language disorder resulting from physical damage to specific brain regions. It may adversely affect multiple communication modalities, including auditory comprehension, verbal expression, reading and writing [1]. In the aspect of verbal expression, the most prevalent deficit among people with aphasia (PWA) is anomia, which refers to word retrieval difficulty [2]. People with dementia or progressive aphasia may also suffer from anomia [3, 4]. Paraphasia is a dominant symptom of anomia. It is characterized by the production of unintended words [5], which comprises a few different forms: (1) phonemic, e.g., the target word "pike" substituted by "pipe"; (2) verbal, e.g., "cat" substituted by "dog"; and (3) neologistic, e.g., target word substituted by a gibberish word. In addition, anomic speech is typified by excessive word-finding pauses [6]. Apart from anomia, symptoms like circumlocution, voice disorder and dysprosody may be present in PWA at various severity level and with different combinations [7, 8].

Speech assessment is considered as an essential component in the clinical process of evaluating the type and severity of aphasia. Automatic speech recognition (ASR) technology has demonstrated great potential in achieving automated analysis and assessment of PWA speech. In [9, 10, 11], impairment-related acoustic and text features derived from ASR output were used to predict subjective assessment scores. It was shown that the ASR accuracy on PWA played a vital role in the feature extraction process. Multiple attempts were made to improve acoustic model (AM) training with domain adaptation [12, 13] and multi-task learning strategies [9, 11]. However, there seems to be a fundamental limitation on the ASR performance for PWA speech, due to the wide variety of unseen phonemic errors and gibberish words. As a matter of fact, these paraphasia symptoms are valuable markers for evaluating severity and type of aphasia [14, 15]. The low ASR accuracy hinders reliable paraphasia detection based on erroneous ASR output [13].

In the present study, we investigate a novel approach of applying ASR to extract speech features that are related to paraphasia, speaking rate and voice abnormality. We propose to regard intractable paraphasias as out-of-vocabulary (OOV) words in ASR. Posterior-based confidence measures of ASR have been widely used for detecting OOVs [16, 17, 18]. In [17], pattern comparison between frame-level phone posteriors generated from concurrent strong recognizer and weak recognizer was found to be useful in detecting OOVs and ASR errors. We extend this approach by stacking frame-level phone posteriors to form a 2-dimensional posteriorgram, i.e., a time-posterior matrix. The discrepancies between the posteriorgrams from a strongly constrained recognizer based on lattice and a weakly constrained one based on AM are exploited to quantify the paraphasia and hence measure the severity of aphasia. In this way, explicit detection of phonemic errors and gibberish words is not required. Apart from paraphasia symptoms, the posteriorgrams contain other impairment-related information, like rate of speaking, pause duration, voice change, and atypical articulation [19].

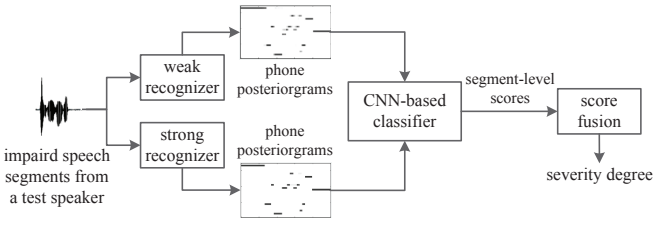


Fig. 1: The proposed system for PWA speech assessment.

As shown in Fig. 1, given an input speech segment from a test speaker, a pair of posteriorgrams is computed from a weak and a strong recognizer. A convolutional neural network (CNN) is trained to classify the posteriorgram pairs and generate a score of severity for the speech segment. We investigate two different CNN models, namely the Siamese [20] and 2-channel CNN [21], which were used for comparing image patches. To obtain speaker-level assessment re-

This research is partially supported by a direct grant from Research Committee of the Chinese University of Hong Kong and a GRF project grant (Ref: CUHK14227216) from the Hong Kong Research Grants Council.

sult, a fusion process is applied to combine all segment-level scores of the same speaker.

2. ASR SYSTEM FOR APHASIA ASSESSMENT

2.1. Corpus

Cantonese AphasiaBank (CanAB) is a large-scale multi-modal corpus jointly developed by the University of Central Florida and the University of Hong Kong [22]. The corpus contains recordings of spontaneous speech from 149 unimpaired and 104 aphasic subjects. The speech recordings were elicited in 8 narrative tasks, including picture description, procedure description, story telling and personal monologue. All impaired subjects participated in a standardized assessment: Cantonese Aphasia Battery [15]. It involves a series of sub-tests measuring speech fluency, naming abilities, etc. The sum of sub-test scores is named the Aphasia Quotient (AQ). The value of AQ (0-100) is an indication of overall severity of impairment. Lower AQ value means higher degree of severity.

About 12.6 hours speech data from 101 unimpaired speakers are used for AM training. The test set contains 1.8 hours speech data from 17 unimpaired speakers and 12 hours speech data from 82 impaired speakers (AQ: 27.0-99.0). The training set and test set are domain- and style-matched.

2.2. ASR System Setup

To mitigate the data scarcity problem for the development of ASR system on impaired speech, we follow the multi-task learning approach as in our previous work [11]. The time-delay layers stacked with bidirectional long short term memory layers (TDNN-BLSTM) are trained with the domain-matched CanAB corpus and two large-scale domain-mismatched Cantonese speech corpora (106.7h) using multi-task learning strategy (MT-TDNN-BLSTM). The domain-matched CanAB is set as the primary task with the highest weight in the loss function, while other two corpora are set as secondary tasks. The phone set contains 32 Cantonese phones (13 vowels and 19 consonants), 1 silence and 1 laughter. A context-dependent GMM-HMM (CD-GMM-HMM) AM is trained beforehand for each task to obtain tied triphone state alignments. Refer to [11] for the details of unimpaired corpora and CD-GMM-HMM AM training. Kaldi [23] is used to train the AMs.

Input Features: 40-dimensional MFCCs without cepstral truncation and 3-dimensional pitch features are extracted from speech audios, with 25ms window length and 10ms window shift. Input features are the concatenation of 100-dimensional i-vectors and 43-dimensional features, where the 43-dimensional features are spliced with context size of 5 frames (2 in past and 2 in future).

MT-TDNN-BLSTM: The TDNN-BLSTM contains 4 TDNN layers (1024 neurons per layer) followed by 4 pairs of forward-backward projected LSTM layers (1024-dimensional cells and 256-dimensional recurrent projections). For the TDNN layers, the number of input contexts used to compute an output activation is $[-2, 2]$ at the 1st layer, $\{0\}$ at the 2nd layer and $[-1, 1]$ at the 3rd and 4th layers. The combined TDNN-BLSTM layers are shared by all tasks. Each task has an independent softmax layer for triphone state prediction.

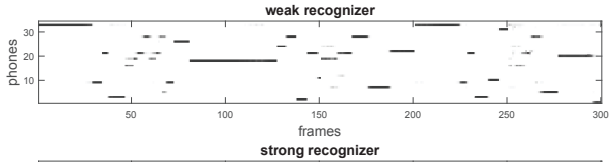
ASR Performance: The automatic transcription is decoded with a syllable bi-gram language model (LM). It is trained with the transcription of training data of CanAB. The system performance, in terms of the syllable error rate (SER), is evaluated on the test set of CanAB. For the 17 unimpaired speakers, the overall SER is 16.73%, while the overall SER for 82 aphasia speakers is 38.20% due to the language impairment. With this ASR system, we extract the phone posteriorgrams based on the following recognizers.

3. POSTERIORGRAMS OF STRONG AND WEAK RECOGNIZERS

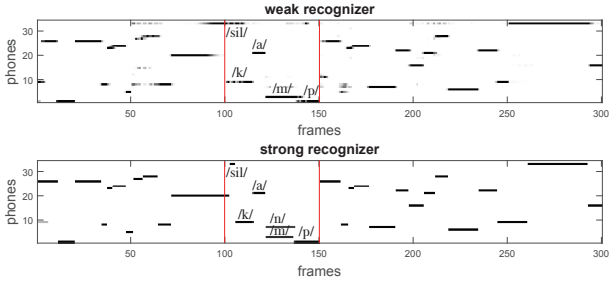
3.1. Weak Recognizer vs. Strong Recognizer

The weak recognizer is a phone posterior estimator based on the MT-TDNN-BLSTM AM. Frame-level phone posterior probabilities are obtained from the softmax layer of the primary task. Each neuron in the softmax layer points to a triphone state. Each of the 34 modeled phones is associated with a set of designated triphone states. The frame-level posterior for the phone is calculated by summing up all the associated state posteriors.

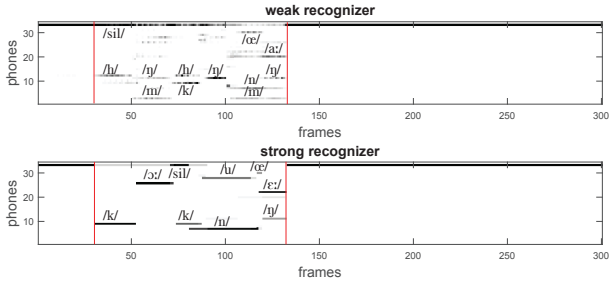
The strong recognizer refers to a lattice-based phone posterior estimator of the primary task. Both the AM and the syllable bi-gram LM are applied such that the decoding output has a higher certainty of predicted phones than the weak recognizer. The frame-level posteriors are derived by applying the Forward-Backward algorithm on the lattice (using the "lattice-to-post" function in Kaldi). The acoustic and language model scaling factors are set as 0.1 and 1.0 respectively.



(a) Posteriorgrams from an unimpaired speech segment.



(b) Posteriorgrams from a mildly impaired speech segment.



(c) Posteriorgrams from a severely impaired speech segment.

Fig. 2: Weak-strong phone posteriorgram pairs from unimpaired, mildly impaired and severely impaired speech segments.

3.2. Patterns in Posteriorgrams

Fig. 2 shows a few examples of weak-strong posteriorgram pairs that represent unimpaired and impaired speech. The posteriorgram pair in Fig. 2(a) is obtained from an unimpaired speech segment. For Fig. 2(b) and 2(c), the speech segments are from a mildly impaired speaker (AQ: 97.2) and a severely impaired speaker (AQ: 42.0), respectively. All speech segments have the same duration of 3 seconds. In the severely impaired case, there is a significant mismatch between the two posteriorgrams, which is related to a gibberish word. Whilst for unimpaired speech, posteriorgrams from the weak and the strong recognizers are highly similar. In general, the weak-strong posteriorgrams of PWA speech show the following characteristics:

- ASR confidence: inconsistency of posterior distribution patterns between the two recognizers; the strong recognizer is confused about predicted phones, i.e., multiple phones activated at the same time frame;
- Voice abnormality: significant perturbation of phone posteriors [19] (level of darkness) given by the weak recognizer;
- Speaking rate: long duration of silence and only a few phones activated;

The above impairment characteristics are more significant in the severely impaired case than in the mild one. This suggests that the weak-strong posteriorgram pairs could be used as input features for detecting and quantifying impairment in PWA speech.

4. EXPERIMENTAL SETUP

4.1. Posteriorgram Features and Feature Labels

The effectiveness of the posteriorgram features is evaluated in a binary classification task. It aims at discriminating PWA with High-AQ (AQ $\geq$ 90) from those with Low-AQ (AQ $<$ 90). The cut-off value of 90 is set to reach balanced number of subjects in two groups. There are 35 subjects in the High-AQ group (label 1) and 47 subjects in the Low-AQ group (label 0). Pairs of posteriorgram features (300×34) are extracted from speech segments of 3 second long. The classification label of each segment is inherited from the impaired speaker. Therefore, the more segments are classified as label 1, the more likely the speaker is in the High-AQ group. There are $4,984$ segments from High-AQ speakers with label 1 and $9,034$ segments from Low-AQ ones with label 0.

The binary classification is carried out by the 5-fold cross validation strategy. In each fold, 80% of the subjects are used for training and the rest 20% subjects are for test. 10% subjects are randomly selected from training subjects as the validation data.

4.2. CNN-based Classifiers

4.2.1. Siamese and 2-channel CNN models

Siamese: The model structure is motivated by the model in [24], as shown in Fig. 3 (left). From the bottom, two CNNs share the identical structure and weights. They are regarded as descriptor computation modules to extract high-level representations from two branches of "weak" and "strong" posteriorgrams. It is followed by a similarity computation layer to capture the pattern mismatch between two posteriorgrams. The formula of the distance computation is $|C(W)-C(S)|$ (element-wise absolute difference), where $C(W)$ and $C(S)$ are representation vectors of "weak" and "strong" posteriorgrams generated from CNNs. A fully-connected layer with the size of 40 is on the top of the model, followed by a sigmoid function to output a final score for the speech segment. The ReLU activation and dropout regularization are applied between the similarity layer and the fully-connected layer.

2-channel: As an extension of siamese model, the 2-channel CNN model shown in Fig. 3 (right) has no explicit module of descriptor [21]. The posteriorgrams from weak and strong recognizers are treated as a 2-channel image that is directly fed into the CNN. In this way, the information contained in two posteriorgrams is jointly processed from the start of the network to predict the severity degree. A 40-dimensional fully-connected layer is applied. The output scores for speech segments are also generated by the sigmoid function.

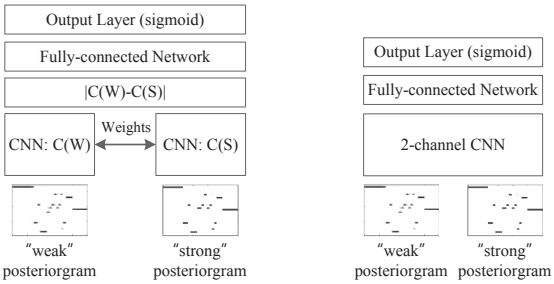


Fig. 3: Architectures of siamese (left) and 2-channel (right) models.

4.2.2. CNN structure variations

In this study, we experiment with two variants of the CNN structures and apply them to siamese and 2-channel models respectively.

Structure 1 (s1): Each row of the posteriorgram (300×34) indicates a frame-level phone posterior. Considering this physical meaning, the structure in [25] for sentence classification is adopted, in which each row of the input matrix denotes a word vector. 5 filter sizes of $\{3, 4, 5, 6, 7\} \times 34$ with stride 1 are applied to the convolutional layer. 20 filters are used per filter size, resulting in 100 feature maps (300×1). It is followed by global average pooling on each feature map and the results are concatenated to a 100-dimensional vector representation. For the *siamese-s1* model, a pair of 100-dimensional vectors denotes $C(W)$ and $C(S)$ derived from two posteriorgrams for the similarity computation. For the *2-channel-s1* model, the 100-dimensional vector is fed to the fully-connected network and further fed to the output layer with dropout regularization.

Structure 2 (s2): This structure is a slight variant of the AlexNet [26]. We revise the first convolution layer to have a kernel size of 11×7 with and a stride of 6×1. Batch normalization, ReLU activation and maxpooling are applied after each convolution layer. For the *siamese-s2* model, a fully-connected layer (3982-dimensional) is applied after the 5 convolutional layers, leading to vector representations of $C(W)$ and $C(S)$. For the *2-channel-s2* model, the sizes of two fully-connected layers are set as 3982 and 40 respectively. Batch normalization and ReLU activation are added after each fully-connected layer. Other settings are consistent with the AlexNet.

4.2.3. Hyperparameters for model training

The training parameters are set empirically. The mini-batch sizes are 64 for training four classifiers: *siamese-s1*, *siamese-s2*, *2-channel-s1*, and *2-channel-s2*. The initial learning rates are set to 1e-3 for siamese models and set to 1e-4 for 2-channel models. Model training aims at minimizing the binary cross-entropy loss with the Adam optimizer (weight decay coefficient 5e-4) [27]. Dropout method with probability 0.5 is used for the regularization purpose. All models are implemented using Pytorch [28].

Table 1: Classification performances of four classifiers using fusion methods by hard decision, averaging and SVM.

Model	Fusion by hard decision			Fusion by averaging			Fusion by SVM		
	Accuracy	F1	Specificity/Recall	Accuracy	F1	Specificity/Recall	Accuracy	F1	Specificity/Recall
Siamese-s1	0.817	0.852	0.915/0.686	0.817	0.854	0.936/0.657	0.829	0.860	0.915/0.714
Siamese-s2	**0.854**	**0.870**	0.851/0.857	0.842	0.857	0.830/0.857	0.866	0.884	0.894/0.829
2-channel-s1	0.817	0.845	0.872/0.743	0.829	0.854	0.872/0.771	0.842	0.857	0.830/0.857
2-channel-s2	**0.854**	**0.870**	0.851/0.857	**0.878**	**0.891**	0.872/0.886	**0.878**	**0.891**	0.872/0.886

5. RESULTS AND DISCUSSION

5.1. Segment-level Classification Accuracy

For the binary classification on test segments from high-AQ and low-AQ subjects, the Area Under receiver operating characteristic Curve (AUC) [29] is used as the performance metric. An AUC value 0.5 means random guess and 1.0 represents perfect classification [30].

Table 2: AUC of test data in each fold based on four CNN classifiers.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Siamese-s1	0.73	0.82	0.70	0.78	0.76
Siamese-s2	0.71	0.81	0.70	0.80	0.78
2-channel-s1	0.72	**0.83**	0.71	0.82	**0.81**
2-channel-s2	**0.74**	**0.83**	**0.73**	**0.83**	**0.81**

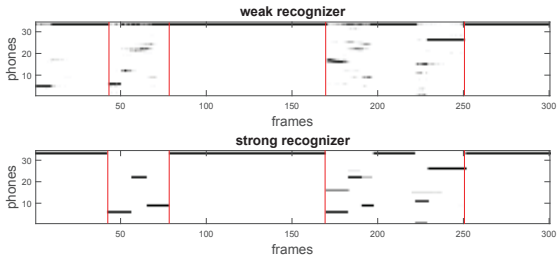


Fig. 4: A speech segment with score 0.15 from *2-channel-s2* model.

Table 2 shows the AUC results of 5-fold cross-validation experiments with different classifier structures. The 2-channel models generally perform better than the siamese models. The 2-channel models allow joint learning of the knowledge from both patches at the first layer [21]. The siamese models put more focus on contrasting the mismatched patterns between weak and strong posteriorgrams, while the 2-channel models are able to learn more comprehensive features, including the mismatched patterns, speaking rate and voice abnormality. For the 2-channel models, the structure s2 performs slightly better than s1. As an example, with the *2-channel-s2* model, a speech segment from an aphasia speaker (AQ: 73.8) is given a predicted score of 0.15. As shown in Fig. 4, the mismatched patterns as well as the low speaking rate lead to the low predicted score. This demonstrates that the CNN is able to learn impairment-related features in posteriorgrams for PWA speech assessment.

5.2. Speaker-level Classification Accuracy

A post-processing method is required to combine all segment-level scores for a test speaker and give a speaker-level classification decision. The following score fusion methods are considered:

1. Fusion by hard decision: The segment-level scores are quantized to the value of 1 (High-AQ) or 0 (Low-AQ) using the threshold of 0.5. If more than 50% of the test segments are with value 1, the speaker is classified as High-AQ, otherwise the speaker is classified as Low-AQ.

2. Fusion by averaging: A speaker-level score is computed by taking the average of segment-level scores. If the speaker-level score is higher than 0.5, the speaker is classified as High-AQ, otherwise he/she is classified as Low-AQ.

3. Fusion by SVM: A 7-dimensional vector of statistical parameters is derived from segment-level scores to represent a speaker. The parameters are: mean, maximum, minimum, standard deviation, $1/4$ quantile, $3/4$ quantile and skewness. A support vector machine (SVM) with polynomial kernel is used to classify the feature vectors. Leave-one-out cross-validation method is adopted.

Table 1 shows the speaker-level binary classification results with different fusion methods. The classification performance is measured in terms of the Accuracy, F1 score, Specificity and Recall. Overall speaking, the *2-channel-s2* model shows the best performance among all models. This is consistent with the AUC results as shown in Table 2. Segment-level scores from the *2-channel-s2* model tend to be more polarized, meaning this model is more certain about target classes. Nevertheless, the *2-channel-s2* model does not benefit from the SVM fusion method for speaker-level classification, whilst the SVM method shows high effectiveness with other models. It is also observed that the CNN structure s2 outperforms the s1 in both siamese and 2-channel models in the speaker-level classification. This suggests that the structure s2 with smaller filter sizes for learning more localized features to assessment is preferred.

In [10], a 5-dimensional feature vector of supra-segmental duration and a 7-dimensional vector of combined features (two syllable embeddings and five duration parameters) were evaluated on the same task of binary classification between High-AQ and Low-AQ subjects. All of the features were obtained from the output of an ASR system, which used a deep neural network based AM and the same LM as the strong recognizer in this study. The F1 scores attained with a random forest classifier were 0.821 for duration features and 0.903 for combined duration and text features. The proposed posteriorgram features perform much better than duration features (0.891 vs. 0.821). This is expected as a posteriorgram contains not only duration information but also paraphasia characteristics. For the combined features, syllable embeddings derived from ASR output, though erroneous, are able to explicitly capture the semantic information of input speech. Such information is not available in the posteriorgram features. It is expected that joint use of these two types of features would lead to further improvement of system performance.

6. CONCLUSIONS

This paper presents a novel approach to automatic assessment of narrative speech from PWA, based on contrasting pairs of phone posteriorgrams from strong and weak recognizers. Impairment-related characteristics in posteriorgrams, namely paraphasias, change of speaking rate, and voice abnormality, are learnt with CNN classification models. For binary classification between mild and severe aphasia, a F1 score of 0.891 could be attained. The 2-channel CNN structure with small filter sizes for learning comprehensive and localized features is found to be most suitable in this application.

7. REFERENCES

[1] J. C. Rosenbek, L. L. LaPointe, and R. T. Wertz, *Aphasia: A clinical approach*. Pro Ed, 1989.

[2] H. Manasco, *Introduction to neurogenic communication disorders*. Jones & Bartlett Publishers, 2017.

[3] W. Webb and R. K. Adler, *Neurology for the Speech-Language Pathologist-E-Book*. Elsevier Health Sciences, 2016.

[4] J. D. Rohrer, W. D. Knight, J. E. Warren, N. C. Fox, M. N. Rossor, and J. D. Warren, "Word-finding difficulty: a clinical analysis of the progressive aphasias," *Brain*, vol. 131, no. 1, pp. 8–38, 2008.

[5] H. Goodglass, E. Kaplan, and B. Barresi, *The assessment of aphasia and related disorders*. Lippincott Williams & Wilkins, 2001.

[6] C. G. Goetz, *Textbook of clinical neurology*. Elsevier Health Sciences, 2007, vol. 355.

[7] Wikipedia contributors, "Aphasia — Wikipedia, the free encyclopedia," 2018, [Online; accessed 10-September-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Aphasia

[8] ——, "Anomic aphasia — Wikipedia, the free encyclopedia," 2018, [Online; accessed 10-September-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Anomic_aphasia

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

[10] 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. ICASSP*, 2018, pp. 5994–5998.

[11] 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*, 2018, pp. 3418–3422.

[12] D. Le and E. M. Provost, "Improving automatic recognition of aphasic speech with Aphasiabank." in *Proc. INTERSPEECH*, 2016, pp. 2681–2685.

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

[14] O. Spreen and A. H. Risser, *Assessment of aphasia*. Oxford University Press, 2003.

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

[16] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," in *Proc. ICASSP*, 2008, pp. 4081–4084.

[17] S. Kombrink, L. Burget, P. Matějka, M. Karafiát, and H. Hermansky, "Posterior-based out of vocabulary word detection in telephone speech," in *Proc. INTERSPEECH*, 2009, pp. 80–83.

[18] D. Wang, S. King, J. Frankel, R. Vipperla, N. Evans, and R. Troncy, "Direct posterior confidence for out-of-vocabulary spoken term detection," *ACM Trans. TOIS*, vol. 30, no. 3, p. 16, 2012.

[19] Y. Liu, T. Lee, P. Ching, T. K. Law, and K. Y. Lee, "Acoustic assessment of disordered voice with continuous speech based on utterance-level ASR posterior features," *Proc. INTERSPEECH*, pp. 2680–2684, 2017.

[20] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.

[21] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. CVPR*, 2015, pp. 4353–4361.

[22] A. P.-H. Kong and S.-P. Law, "The Cantonese Aphasiabank," [Online; accessed 10-September-2018]. [Online]. Available: http://www.speech.hku.hk/caphbank/search/

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, no. EPFL-CONF-192584, 2011.

[24] 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," *Proc. INTERSPEECH*, pp. 1427–1431, 2017.

[25] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[29] M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," *Metodoloski zvezki*, vol. 3, no. 1, p. 89, 2006.

[30] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.