INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automatic Assessment of Language Impairment Based on Raw ASR Output

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

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

[1]yingqin@link.cuhk.edu.hk, [1]tanlee@cuhk.edu.hk, [2]antkong@ucf.edu

Abstract

For automatic assessment of language impairment in natural speech, properly designed text-based features are needed. The feature design relies on experts' domain knowledge and the feature extraction process may undesirably involve manual effort on transcribing. This paper describes a novel approach to automatic assessment of language impairment in narrative speech of people with aphasia (PWA), without explicit knowledge-driven feature design. A convolutional neural network (CNN) is used to extract language impairment related text features from the output of an automatic speech recognition (ASR) system or, if available, the manual transcription of input speech. To mitigate the adverse effect of ASR errors, confusion network is adopted to improve the robustness of embedding representation of ASR output. The proposed approach is evaluated on the task of discriminating severe PWA from mild PWA based on Cantonese narrative speech. Experimental results confirm the effectiveness of automatically learned text features. It is also shown that CNN models trained with text input and acoustic features are complementary to each other.

Index Terms: speech assessment, language impairment, aphasia, CNN

1. Introduction

Aphasia refers to an acquired communication disorder due to focal brain damage, which impairs a person's ability to process language. It is most commonly caused by a stroke. Multiple aspects of the language system, such as phonology, lexicon, syntax and semantics, can be adversely affected by Aphasia [1]. People with Aphasia (PWA) may show various types of symptoms, e. g. inability to pronounce, difficulty in naming objects, forming words and/or comprehending language [2]. Speech assessment is an important component of the comprehensive assessment system for PWA. It is typically carried out by trained speech and language pathologists as a clinical process, aiming to determine the severity and/or type of impairment. Subjective assessment of language impairment requires not only clinical knowledge about the disease but also similar language and cultural backgrounds between the subject and the clinician. The shortage of desired pathologists and time-consuming subjective assessment system restrict PWA to the assessment regularly.

Automatic assessment of PWA speech has been investigated to achieve low-cost objective diagnosis of PWA. The main focuses are on the acoustic aspect and the linguistic aspect of speech abnormality. Fraser et al. investigated sub-type classification on primary progressive aphasia using a large set of acoustic and text features. The features were either manually measured by trained persons [3, 4] or automatically computed with a commercial Automatic Speech Recognition (ASR) system [5]. However, the low ASR accuracy on PWA speech made this approach infeasible in clinical use. In [6, 7], Duc Le et al. attempted to improve ASR performance on impaired speech by applying discriminative pre-training and multi-task techniques. A combined set of clinically-relevant acoustic and text features derived from ASR outputs were utilized to predict subjective assessment scores. In our previous work, syllable-level embedding features, phone posteriorgrams, and supra-segmental duration features, were shown to be effective in assessment of language impairment with Cantonese-speaking PWA [8, 9, 10].

Generally speaking, use of expert knowledge in feature design and extraction played an important role in most previous studies on automatic assessment of language impairment. Recently, we attempted an end-to-end approach to the problem, in which a direct mapping from conventional acoustic features (i.e., Log-Mel filter-bank features) to assessment score was realized with a Convolutional Neural Network (CNN) model [11]. By visualizing the learned features using class activation mapping [12], it was confirmed that the CNN model could capture impairment-related acoustic characteristics that are in agreement with human-designed features. Nevertheless, the model was not able to learn the linguistic features that are critical to language impairment.

In the present study, we take a further step towards a more comprehensive assessment system that can automatically learn both acoustic and linguistic features. CNN models have been widely applied to the text-based classification [13] and spoken language understanding problems [14, 15, 16]. This motivates us to apply a CNN model to assess PWA speech in the linguistic aspect. The CNN model is trained with manual transcriptions of impaired speech in order to capture language impairment related features from raw text data. The test data can be ASR output or manual transcriptions of impaired story-telling speech. To mitigate the effect of ASR errors on the assessment performance, confusion networks are incorporated to the embedding representation of ASR output. Lastly, we compare the assessment results between CNN models trained with text data and acoustic features. It is found that the results can be complementary to each other, such that joint training of two CNN models to perform acoustic analysis and linguistic analysis on PWA speech is promising.

2. Database

Cantonese AphasiaBank (CanAB) is a large-scale multi-modal database developed by University of Central Florida and The University of Hong Kong [17], which contains recordings of spontaneous speech from 104 aphasic subjects and 149 unimpaired subjects. All subjects were required to complete 8 narrative tasks, including picture description, procedure description, story telling and personal monologue. Except the personal monologue, the remaining 7 tasks are about specific topics. Each of them is referred as a "story". The speech recordings were manually transcribed using the CLAN program [18].

http://dx.doi.org/10.21437/Interspeech.2019-1688

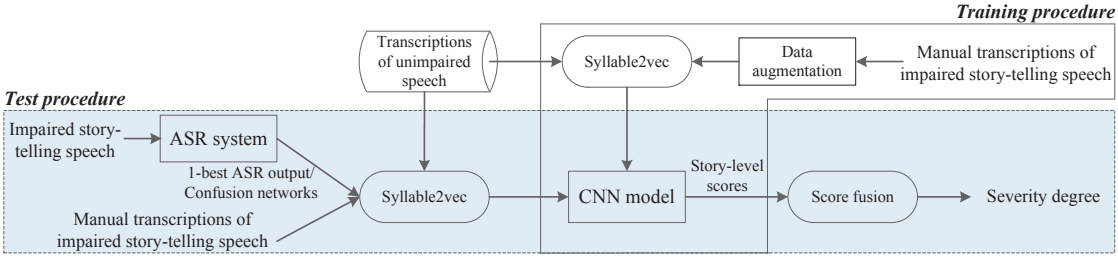

Figure 1: *General framework of the proposed assessment system. The CNN model is trained with manual transcriptions of narrative speech from PWA and tested with ASR output or manual transcriptions of story-telling impaired speech.*

The orthographic transcriptions are in the form of sequences of Chinese characters and can be converted into syllable transcriptions using a pronunciation lexicon [19]. Each story in CanAB was manually segmented into sentences. In this study, speech from 118 unimpaired speakers (control group) with 8 tasks (14.4 hours) and speech from 92 impaired speakers (test group) with 7 tasks (13.1 hours) are used for the following experiments. Table 1 shows detailed descriptions of 7 tasks and word distribution of each task, i. e. average and standard deviation (std) for 92 impaired speakers.

All impaired subjects were requested to go through a standardized assessment system named Cantonese Aphasia Battery [20]. It involves a set of sub-tests measuring speech fluency, naming abilities, etc. The sum of sub-test scores is an indication of overall severity of impairment, which is known as the Aphasia Quotient (AQ). Its value ranges from 0 to 100. Lower AQ value indicates a higher severity degree.

Table 1: *Descriptions of* 7 *tasks and word distribution (average and std) of each task for* 92 *impaired speakers in CanAB.*

Content of 7 stories in CanAB (# of stories)	Average # of words (std)
A boy accidentally breaks a window (92)	70.6 (44.1)
A cat on a tree being rescued (92)	98.4 (77.9)
Fairy tale titled 'The boy who cried wolf' (92)	139.5 (104.2)
Prepare a sandwich with egg, ham and bread (92)	58.6 (46.2)
A fireman rescuing a girl (92)	67.4 (64.8)
A boy refuses an umbrella from his mother (92)	83.8 (54.1)
Fairy tale titled 'The tortoise and the hare' (92)	143.5 (120.6)

3. Methods

3.1. General framework

Figure 1 illustrates the general framework of our proposed assessment system. It aims at discriminating PWA with High-AQ (AQ $\geq$ 90) from those with Low-AQ (AQ $<$ 90). The system is trained with syllable-level manual transcriptions of impaired narrative speech. The training procedure contains the components of data augmentation, generation of syllable-level embedding representations and CNN model training. For the test procedure, the input to the system can be ASR decoding output or manual transcriptions of impaired story-telling speech. With the trained Syllable2vec system, ASR output or manual transcriptions are converted to syllable-level vectors and then fed to the CNN classifier. The output of CNN indicates an assessment score for the input story. To obtain an overall assessment score for each impaired speaker, his/her story-level assessment scores are combined by a score fusion system.

3.2. Data augmentation and syllable embeddings

Data augmentation strategy has been applied to speech recognition [21] and text-based assessment for children with autism spectrum disorder [22]. It helps to increase the quantity of data for training neural networks and improve robustness of the models. Due to limited story-telling resources in the CanAB, a data augmentation is carried out following the method in [22]. A sliding window with stride 1 is performed on a training story with n segmented sentences, leading to another training sample with n sentences. The widths of sliding windows are empirically set to $\{3, 5, 7\}$ sentences. The final training set contains full-length stories and 3-fold augmented text data.

The Word2vec Toolkit [23] is adopted to implement a continuous bag-of-words (CBOW) model [24] for generating syllable-level embeddings. The non-tonal syllable transcriptions of speech data from CanAB's control group are used to train a Syllable2vec system. It contains about $183,000$ syllables and covers 523 unique items. The number of context words for training CBOW model is set to 6. The dimension of word vectors is set to 50. Each training sample for CNN is represented by a concatenation of syllable-level embeddings and then zero-padded to the maximum length (794) among all stories.

3.3. CNN model for classification

The implementation of CNN model follows the idea in [13]. 2-dimensional representation of an input story $\boldsymbol{X}$ with the shape of $l \times d$ is fed into the convolutional layer, where $l = 794$ and $d = 50$ in this study. Let $\boldsymbol{x}_{i:i+j}$ denotes the concatenation of syllable vectors $\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_{i+j}$. After passing through a filter $\boldsymbol{w}$ with the size of $h \times d$ in the convolutional layer, a feature m_i is generated based on a window of h syllables:

$$m_i = f(\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h-1} + b), \tag{1}$$

where $f(\cdot)$ is a ReLU function and b is a bias term. The filter is moved with stride 1, thus a feature map $\boldsymbol{m}$ can be obtained by

$$\boldsymbol{m} = [m_1, m_2, \cdots, m_{l-h+1}]. \tag{2}$$

In the present study, 4 filter sizes of $\{2, 3, 4, 5\} \times 50$ which correspond to N-gram orders (i. e. from bi-grams to fifth-grams) are utilized. There are 100 filters for each type of filter size, resulting in 400 feature maps. Max pooling is performed on each feature map, which is expected to capture the most important feature from each feature map and naturally tackle the variable sentence lengths [13]. The results are concatenated to a 400-dimensional vector. This vector representation forms the penultimate layer of the CNN model. It is passed to a fully-connected sigmoid layer with dropout regularization to output an assessment score for each input text sample.

3.4. Assessment based on ASR output

3.4.1. ASR system

The development of ASR system on impaired speech follows multi-task learning approach in our previous work [9]. Time-delay neural network combined with bi-directional long short-term memory layers (TDNN-BLSTM) are shared by three phone-level acoustic modeling tasks. The primary task is trained with speech from control group of CanAB for modeling the spontaneous style speech. The other two tasks are trained with two large-scale domain-mismatched Cantonese speech corpora (106.7 hours), which target to model read style speech. The softmax layer is independently assigned to each task to predict tri-phone states. A context-dependent GMM-HMM (CD-GMM-HMM) for each task is trained beforehand to generate state-level tri-phone alignments. Refer to [9] for the detailed information of training corpora and CD-GMM-HMM training.

For each frame of speech, 40-dimensional Mel-frequency cepstral coefficients and 3-dimensional pitch features are extracted. Input features to the multi-task TDNN-BLSTM are the concatenation of these 43-dimensional features and 100-dimensional i-vectors, where the 43-dimensional features are spliced with a context size of 5 frames (2 in the past and 2 in the future: $[-2, 2]$). The structure of TDNN-BLSTM consists of 4 TDNN layers (1024 neurons per layer) stacked with 4 pairs of forward-backward projected LSTM layers containing 1024-dimensional cells and 256-dimensional recurrent projections per layer. For the TDNN layers, a context window of input frames to compute an output activation is $\{0\}$ at the 2nd layer and $[-1, 1]$ at the 3rd and 4th layers.

For ASR decoding, the language models are syllable bi-grams trained with orthographic transcriptions of speech from control group in CanAB. The ASR performance on 92 impaired speakers is evaluated in terms of syllable error rate (SER), which achieves optimal SER of 39.4%.

3.4.2. Syllable embeddings with confusion networks

Confusion networks (CNs) are compact graphical representations of ASR hypotheses in the lattice [25]. The CN aligns a set of candidate syllables at each position with their associated posterior probabilities, which is represented by a sequence of bag-of-weighted-arcs. Our previous study [9] demonstrated that the CNs were useful to improve the robustness of text features to ASR errors for the assessment of PWA. In addition, CNs were adopted to train a BLSTM-RNN system for spoken utterance classification and showed better classification performance than using the 1-best ASR output [26]. Motivated by the approach in [26], we propose to incorporate CNs into the syllable embeddings as the input to CNN in this study.

Let L refers to the length of position segments in a CN, and $N_1, N_2, \cdots, N_L$ represent the number of candidate syllables for all positions. For the l^{th} segment, the candidate syllables are $w_{1,l}, w_{2,l}, \cdots, w_{N_l,l}$ with the posterior probabilities $p_{1,l}, p_{2,l}, \cdots, p_{N_l,l}$. Incorporated with the CN, a modified weighted-sum syllable vector representation $\boldsymbol{x}_{\boldsymbol{l}}^{\text{modified}}$ at the l^{th} position is given by

$$\boldsymbol{x}_{\boldsymbol{l}}^{\text{modified}} = \sum_{i=1}^{N_l} p_{i,l} \boldsymbol{x}_{i,l}, \tag{3}$$

where $\boldsymbol{x}_{i,l}$ denotes the syllable vector of $w_{i,l}$. The modified syllable vectors of a text sample are further concatenated as a 2-dimensional representation to the CNN model.

3.5. Score fusion

A post-processing procedure is required to fuse all story-level scores of a test speaker as an overall assessment decision. We compute 8 statistical parameters based on 7 story-level scores per impaired subject, including mean, maximum, minimum, standard deviation, $1/4$ quantile, $3/4$ quantile, skewness and kurtosis. They are formed as an 8-dimensional feature vector for each impaired subject. The score fusion is carried out using a Support Vector Machine (SVM) with radial basis function kernel. The leave-one-out cross validation approach is adopted. Feature vectors from 91 impaired speakers are used to train the SVM model and the rest one is as the test data. We use grid search strategy to determine the best parameter settings of SVM based on the criterion of classification accuracy.

4. Experimental setup

4.1. Setup for training CNN model

The binary classification experiment (High-AQ vs. Low-AQ) is carried out using 5-fold cross validation strategy. In each fold, text data from 80% subjects are used for training and those from the rest 20% subjects are used for test. We randomly select 10% subjects from training subjects to form a validation set. There are 39 PWA in High-AQ group with classification label 1 and 53 PWA in Low-AQ group with label 0. The classification labels of input text samples are inherited from the impaired speaker.

The hyperparameters for training the CNN model are chosen empirically. The mini-batch size is set to 64 and the initial learning rate is set to 10^{-3}. Binary cross-entropy is used as the loss function, which is optimized using the Adam algorithm with weight decay coefficient 5×10^{-4} [27]. Dropout technique with probability 0.5 is used for the regularization purpose. The CNN model is implemented using the Pytorch Toolkit [28].

4.2. Baseline systems

The proposed system is compared to the following baseline systems with known manual transcriptions of impaired speech:

1. Perplexity of N-grams [29]: $\{2, 3, 4, 5\}$-grams language models are trained with manual transcriptions of speech from CanAB's control group. They are applied to the transcriptions of 7 types of spoken stories from 92 impaired subjects to compute perplexity values. For each impaired speaker, we take the average of perplexities given by 4 types of N-grams, resulting in a 7-dimensional vector. Note that this approach is no need to follow the 5-fold cross validation scheme.

2. Average of syllable vectors (Average-syllvec) [8]: Given an impaired story, a 50-dimensional story-level embedding is obtained by averaging all syllable vectors (see section 3.2) in accordance to the transcription of the story. In each fold, a logistic regression model is trained with the story-level embeddings of training set to output assessment scores for 7 stories per test speaker.

3. Doc2vec [30]: The impaired stories are represented as vector representations using the Doc2vec technique. Each 50-dimensional story-level vector is a concatenation of two vectors: one learned by distributed memory model with the dimension of 30 and the other one learned by distributed bag-of-words model with the dimension of 20. The Doc2vec is also trained with manual transcriptions of speech from CanAB's control group. Similar to previous case, 7 assessment scores for each impaired

speaker are obtained based on logistic regression model.

After obtaining 7-dimensional score vectors for impaired speakers, the score fusion system mentioned in section 3.5 is applied to make final speaker-level classification decisions for above baseline systems.

5. Results and discussion

5.1. Classification performance

5.1.1. Story-level classification accuracy

Before passing to the score fusion model, the direct classification performance of models can be reflected by story-level classification accuracy. The performance metric we use is the Area Under receiver operating characteristic Curve (AUC). An AUC value of 0.5 represents a random guess and 1.0 means a perfect classification. Except the first baseline system, we compare the AUC results of 5-fold experiments using CNN model and other baseline systems tested with manual transcriptions. The performance of CNNs tested with 1-best ASR output and CNs is also compared. Table 2 summarizes the AUC results.

In the case of using manual transcriptions as test data, the best AUC result is obtained by the CNN model. It significantly outperforms both baseline methods of the average of syllable vectors and Doc2vec, probably because the CNN is specialized in capturing more localized language impairment related features from text data. Due to the ASR errors, an obvious AUC reduction can be seen when the type of test data changes to 1-best ASR output. By incorporating rich ASR hypotheses in CNs into the syllable embedding representations, the AUC performance is slightly improved. This demonstrates that using this type of modified syllable embeddings as the input to CNN is able to mitigate the effect of ASR error on the assessment performance.

Table 2: *AUC results of test data in 5 folds. We compare the performance of CNN model and two baseline systems tested with manual transcriptions of impaired speech. The performance of CNN model tested with 1-best ASR output and CNs is also compared.*

Data type for test	Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Manual transcriptions	Average-syllvec	0.64	0.51	0.71	0.77	0.75
Manual transcriptions	Doc2vec	0.56	0.57	0.64	0.74	0.71
Manual transcriptions	CNN	**0.79**	**0.73**	**0.82**	**0.80**	**0.90**
1-best ASR output	CNN	0.73	**0.62**	**0.81**	**0.79**	0.79
Confusion networks	CNN	**0.75**	**0.62**	0.78	**0.79**	**0.82**

5.1.2. Speaker-level classification accuracy

The first part of Table 3 compares the speaker-level classification performance of proposed CNN model and all baseline systems with manual transcriptions as test data. The results in terms of accuracy and average F1 score are given by the SVM fusion model. It is shown that the CNN model performs better than all baseline systems. One possible cause is that N-grams and semantic content related features in text data can be jointly captured by CNN. Another important reason is that the supervised CNN model is able to jointly optimize the parameters of feature extraction and classification models.

Manual transcriptions can be regarded as the output from an ideal ASR system, such that the classification accuracy of 0.837 is an upper bound of proposed assessment system. When the CNN is tested with 1-best ASR output, the speaker-level classification accuracy significantly decreases. This suggests that the ASR performance on impaired speech should be further improved to narrow the performance gap. With the help of CNs, the assessment accuracy can achieve 0.815, which is much closer to the upper bound performance.

Overall speaking, the CNN model is feasible to perform automatic assessment of language impairment. The CNs is effective to improve robustness of embedding representations of ASR output and thus improving the assessment performance.

Table 3: *Comparison of speaker-level classification performance in terms of accuracy and average F1 score based on manual transcriptions, ASR output and acoustic features.*

Data type for test	Method	Accuracy/F1
Manual transcriptions	Perplexity of N-grams	0.663/0.636
Manual transcriptions	Average-syllvec	0.794/0.786
Manual transcriptions	Doc2vec	0.783/0.778
Manual transcriptions	CNN	**0.837/0.833**
1-best ASR output	CNN	0.761/0.759
Confusion networks	CNN	**0.815/0.809**
Log-Mel acoustic features	CNN	0.826/0.822

5.2. Compare to CNN trained with acoustic features

We follow the approach in [11] to develop an end-to-end assessment system. With the same arrangement of 5-fold cross validation on 92 impaired speakers, a CNN model with global average pooling is trained with Log-Mel acoustic features from PWA speech. The binary classification result is shown in Table 3. It attains comparable performance to the CNN model trained with manual transcriptions, but mainly focus on leaning impairment related features from PWA speech in the acoustic aspect. There are 15 and 16 impaired subjects being mis-classified by manual transcription-based CNN and acoustic feature-based CNN, and only 5 of them are commonly mis-classified. We further analyze one typical test speaker (AQ: 66.8) who is mis-classified to High-AQ group based on acoustic feature-based CNN due to his high speaking rate. However, his spoken story contains very few topic-specific words. This phenomenon is captured by the manual transcription-based CNN such that he is correctly classified as Low-AQ. This reveals the assessment results from these two models can be complementary to each other. A combination of CNN models to perform acoustic and linguistic analyses on impaired speech will be investigated in the following study.

6. Conclusions

In this paper, we present an automatic speech assessment based on narrative speech from Cantonese-speaking PWA. A CNN model is utilized to implicitly and efficiently extract language impairment related text features from ASR output and manual transcriptions of impaired speech. The confusion networks are adopted to mitigate the effect of ASR errors and show feasibility in improving the assessment performance. In the future, a more comprehensive neural network structure capturing both acoustic and linguistic features in PWA speech will be investigated.

7. Acknowledgements

This research is partially supported by the GRF project grants (Ref: CUHK14204014 and CUHK14227216) from the Hong Kong Research Grants Council, the Major Program of National Social Science Fund of China (Ref: 13&ZD189), and the CUHK Shenzhen Research Institute.

8. References

[1] H. Adam, "Dysprosody in aphasia: An acoustic analysis evidence from palestinian arabic," *Journal of Language and Linguistic Studies*, vol. 10, no. 1, pp. 153–162, 2014.

[2] C. Code, *The characteristics of aphasia*. CRC Press, 1989.

[3] K. C. Fraser, F. Rudzicz, and E. Rochon, "Using text and acoustic features to diagnose progressive aphasia and its subtypes." in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 2177–2181.

[4] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *cortex*, vol. 55, pp. 43–60, 2014.

[5] K. C. Fraser, F. Rudzicz, N. L. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proceedings of the 4th Workshop on Speech and Language Processing for Assistive Technologies*, 2013, pp. 47–54.

[6] D. Le and E. M. Provost, "Improving automatic recognition of aphasic speech with Aphasiabank." in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 2681–2685.

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

[8] 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 *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5994–5998.

[9] 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 *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3418–3422.

[10] Y. Qin, T. Lee, and A. P. H. Kong, "Combining phone posteriorgrams from strong and weak recognizers for automatic speech assessment of people with aphasia," in *to appear in Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[11] 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," *arXiv preprint arXiv:1904.00361*, 2019.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

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

[14] P. Xu and R. Sarikaya, "Convolutional neural network based triangular crf for joint intent detection and slot filling," in *Proceedings of Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 78–83.

[15] N. T. Vu, "Sequential convolutional neural networks for slot filling in spoken language understanding," *arXiv preprint arXiv:1606.07783*, 2016.

[16] A. Celikyilmaz, R. Sarikaya, D. Hakkani-Tür, X. Liu, N. Ramesh, and G. Tür, "A new pre-training method for training deep learning models with application to spoken language understanding." in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 3255–3259.

[17] 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," *Behavior research methods*, pp. 1–14, 2018.

[18] B. MacWhinney, *The CHILDES project: The database*. Psychology Press, 2000, vol. 2.

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

[20] 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.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3586–3589.

[22] Y.-S. Liu, C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Learning lexical coherence representation using LSTM forget gate for children with autism spectrum disorder during story-telling," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6029–6033.

[23] Google Inc., "word2vec," accessed 10 December 2018. [Online]. Available: https://code.google.com/archive/p/word2vec/

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[25] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proceedings of EUROSPEECH*, 1999, pp. 495–498.

[26] 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 *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6039–6043.

[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 *Conference on Neural Information Processing Systems Workshop (NIPS-W)*, 2017.

[29] S. Wankerl, E. Nöth, and E. Stefan, "An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language," in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3162–3166.

[30] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of International conference on machine learning (ICML)*, 2014, pp. 1188–1196.