Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning

Ying Qin[1], *Tan Lee*[1], *Siyuan Feng*[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
{yingqin,siyuanfeng}@link.cuhk.edu.hk, tanlee@cuhk.edu.hk, antkong@ucf.edu

Abstract

This paper describes an investigation on automatic speech assessment for people with aphasia (PWA) using a DNN based automatic speech recognition (ASR) system. The main problems being addressed are the lack of training speech in the intended application domain and the relevant degradation of ASR performance for impaired speech of PWA. We adopt the TDNN-BLSTM structure for acoustic modeling and apply the technique of multi-task learning with large amount of domain-mismatched data. This leads to a significant improvement on the recognition accuracy, as compared with a conventional single-task learning DNN system. To facilitate the extraction of robust text features for quantifying language impairment in PWA speech, we propose to incorporate N-best hypotheses and confusion network representation of the ASR output. The severity of impairment is predicted from text features and supra-segmental duration features using different regression models. Experimental results show a high correlation of 0.842 between the predicted severity level and the subjective Aphasia Quotient score.

Index Terms: Speech assessment, aphasia, TDNN-BLSTM, multi-task learning

1. Introduction

Aphasia refers to a type of acquired language impairment typically caused by stroke. The impairment could span across various levels and components of the language system, including phonology, lexicon, syntax, and semantics [1]. Speech assessment is an essential part of the comprehensive assessment process, which mainly aims at determining the type and/or severity of impairment for people with aphasia (PWA). It is realized by acoustical and linguistic analysis of PWA speech elicited through story-telling tasks. There have been a few studies on automatic analysis of speech from PWA. In [2, 3], Fraser et al. used acoustic and text features for automatic classification of sub-types of primary progressive aphasia. The proposed text features were derived from manually prepared speech transcription. In a subsequent study [4], a commercial automatic speech recognition (ASR) system was used to generate the speech transcription for text feature extraction. However, the low recognition accuracy of this general-purpose ASR system on PWA speech limits its practical use in the PWA speech assessment.

In our previous study [5], a framework of fully automatic speech assessment for Cantonese-speaking PWA was developed. A domain-matched ASR system trained with unimpaired speech was used to decode PWA speech into syllable sequences with time alignment information. Supra-segmental duration features were computed from the time alignment, while text features were extracted by a novel method of syllable-level embedding based on the ASR output. The assessment of severity of speech impairment was formulated as a regression problem with the combined features. It was noted that the ASR system had a low accuracy on impaired speech (average syllable error rate (SER) of 48.08%). This inevitably affected the reliability and robustness of the extracted features.

A common problem in developing of ASR systems for atypical speech, including the PWA speech, is due to the lack of training data that are appropriate in terms of spoken content, speaking style, etc. In [5, 6], although a limited amount of domain-matched unimpaired speech were available, they were not sufficient to support the use of most advanced deep learning techniques. In [7] and [8], tandem feature and discriminative pretraining based out-of-domain adaptation methods were applied to improve the ASR performance on a smaller impaired speech corpus. The out-of-domain data could be healthy speech or other types of impaired speech.

In this paper, we propose to use multi-task learning (MTL) strategy to improve the performance of a domain-matched ASR for PWA speech assessment. Under the MTL framework [9], related tasks could be used to share internal representations and jointly learned to improve the generalization of acoustic models [10]. We make use of two large-vocabulary databases of unimpaired Cantonese speech to formulate two auxiliary learning tasks of phone-level acoustic modeling, with the goal of boosting the recognition accuracy for story-telling speech as the main task. Time delay neural network (TDNN) [11] and long-short-term memory recurrent neural network (LSTM-RNN) [12] are well known of the capability of capturing long-term temporal dependency of acoustic events. Bidirectional LSTM (BLSTM) [13] is an extension to LSTM with both preceding and succeeding contexts considered. TDNN-BLSTM has demonstrated its effectiveness in DNN-HMM hybrid acoustic modeling for ASR [10, 14]. This motivates the use of TDNN-BLSTM with MTL in this study.

On the other hand, rich representation of ASR output, such as N-best hypotheses [15] and confusion networks [16], has been widely used in spoken language understanding and speech translation [17, 18]. In order to mitigate the effect of ASR errors, we propose to incorporate rich representation of ASR output into the computation of text features.

2. Datasets

2.1. Domain-matched dataset: Cantonese AphasiaBank

Cantonese AphasiaBank (CanAB) is a large-scale multi-modal database jointly developed by the University of Central Florida and the University of Hong Kong [19]. The corpus contains audio recordings of spontaneous speech from 104 aphasic subjects and 149 unimpaired subjects. Each subject was requested to complete 8 narrative tasks, including 4 picture descriptions,

10.21437/Interspeech.2018-1630

1 procedure description, 2 story telling and 1 personal monologue. Except the personal monologue, the speech produced in each task is expected to be with a specific topic (referred as a "story"). Speech data were manually transcribed into Chinese characters. Fillers, unintelligible speech and non-speech sounds were represented by special symbols. For the development of Cantonese ASR system, the characters were converted into Cantonese syllables using a pronunciation lexicon [20]. All aphasic subjects went through a standardized assessment system using the Cantonese Aphasia Battery [21]. It involves a number of sub-tests measuring the subject's speech fluency, information content and so on [21]. The assessment result is a composite score named the Aphasia Quotient (AQ). The value of AQ (0 - 100) is regarded as an indication of overall severity of language impairment. Lower AQ value implies higher degree of severity.

For various non-technical reasons, not all of the recorded speech in CanAB were accurately transcribed. In this study, 101 unimpaired speakers' speech recordings of 8 tasks (about 12.6 hours) are selected as training set. The test set contains about 10.1 hours speech recordings of 7 tasks (except personal monologue) from 82 impaired speakers (AQ: 27.0 - 99.0), including 52 Anomic, 6 Transcortical sensory, 12 Transcortical motor, 8 Broca's, 1 Isolation, 2 Wernicke's and 1 Global aphasia. The training set and test set are domain- and style-matched.

2.2. Domain-mismatched datasets: CUSENT & K086

CUSENT is a large-scale read speech corpus of Cantonese. It was developed by The Chinese University of Hong Kong [22]. The speech content consists of 5, 100 distinct sentences selected from newspaper articles. There are 20, 378 training utterances from 34 male and 34 female speakers, and 799 test utterances from 4 male and 4 female speakers. The durations for training set and test set are 19.3 hours and 0.6 hours respectively.

King-ASR-086 (K086) [23] is a commercial Cantonese speech database, which contains read speech recordings of 87.4 hours from 136 native Cantonese speakers. The content covers sports, science, international news, etc. 32, 264 utterances (about 71 hours) from 55 male and 55 female speakers are selected as training set, and 7, 754 utterances from 13 male and 13 female speakers (about 16 hours) are selected as test set.

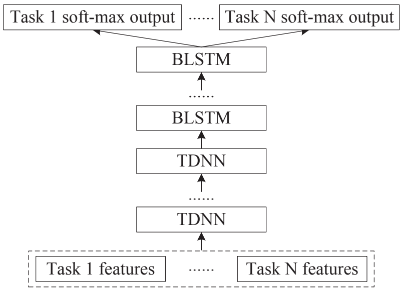

Figure 1: *The architecture of proposed MT-TDNN-BLSTM.*

3. ASR system for aphasia assessment

3.1. MT-TDNN-BLSTM model

In our previous study [5], a standard DNN based ASR for assessment was trained with limited domain-matched healthy speech. MTL provides a potential way to use the domain-mismatched datasets to tackle the data scarcity problem. Furthermore, the advanced TDNN-BLSTM acoustic model shows good performances in large vocabulary speech recognition systems in recent years. Hence, we propose to establish a MT-TDNN-BLSTM based ASR system for PWA speech, as illustrated in Figure 1. The combined layers of TDNN-BLSTM are shared among multiple tasks. During the training procedure, parameters of shared hidden layers and the specific soft-max output layer corresponding to a certain task are updated. The total cross-entropy loss function are weighted across tasks. Refer to [24] for detailed explanation of the MTL.

3.2. System Setup and Performance

Like Mandarin, Cantonese is a monosyllabic and tonal language. Each Chinese character is spoken as a monosyllable with a specific tone. There are a total of 13 vowels and 19 consonants in Cantonese, from which over 600 legitimate base syllables can be formed [20]. Each of 32 phonemes is represented by a hidden Markov model (HMM) with 3 emission states. Kaldi [25] is used to train all acoustic models in this study.

A GMM-HMM acoustic model is trained beforehand for each task to obtain state level phone alignments. The training set for each task is detailed in section 2. The input features used to train GMM-HMMs are 40-dimensional fMLLR features transformed from 39-dimensional MFCC+Δ+$\Delta\Delta$ with a contextual window of 7 frames. In addition to 32 basic acoustic units, five of the most common non-content sounds appeared in CanAB (i.e. inserted filler words, lengthened/repeated initial consonants, para-verbal sounds like laughing and sighing) are modeled by dedicated HMMs and included as part of the acoustic models. Subsequently, we build neural network based ASR systems as follows:

DNN_baseline [5]: A standard feed-forward DNN based ASR system with 6 hidden layers and 1024 neurons per hidden layer is trained to estimate the posterior probabilities of triphone states. The acoustic features are 440-dimensional fMLLR features with a contextual window of 11 frames. The data used to train this system is domain-matched training set of the CanAB corpus.

MT-TDNN-BLSTM: The TDNN-BLSTM model consists of 4 TDNN layers with 1024 neurons per layer followed by 4 pairs of forward-backward projected LSTM layers with 1024-dimensional cells and 256-dimensional recurrent projections. For the TDNN layers, at each time step, the number of input contexts required to compute an output activation are $[-2, 2]$ at the first layer, $\{0\}$ at the second layer, $[-1, 1]$ at third and forth layers. They are all implemented with ReLU and batchnormalization [26]. The combined TDNN-BLSTM layers are shared among three tasks of phone-level acoustic modeling trained with training sets of CanAB, K086 and CUSENT. Each task has an independent soft-max layer for triphone state classification. The domain-matched CanAB is set as the primary task with the highest weight in the loss function, while K086 and CUSENT are set as the secondary tasks. Speech-perturbation method is applied to augment training data three-fold, with speed factor of 0.9, 1.0 and 1.1 [27]. The 40-dimensional MFCCs without cepstral truncation are computed as the input to the neural network [28]. Pitch features have been shown useful to boost ASR performance on tonal language like Cantonese [29]. In addition, i-vector based neural network adaptation has been proved to benefit the ASR on PWA speech [8]. Therefore, MFCCs are further appended with 3-dimensional pitch features designed in [29] and 100-dimensional i-vectors. The appended acoustic features are anticipated to capture pitch information of Cantonese and handle the high speaker variability appeared in impaired speech. The training procedure follows exponential-decay learning schedule from 1.5E$-$3 to 1.5E$-$4. The mini-batch size is 64 and the number of training epochs is 4. Dropout with the probability 0.1 is applied to improve generalization of neural networks [30].

The performances of proposed systems in terms of SER are evaluated on the CanAB test set (PWA speech). The language models are syllable bi-grams trained with the orthographic transcriptions of all training data of CanAB.

Table 1: *Performances of DNN_baseline system and MT-TDNN-BLSTM system evaluated on the CanAB test set.*

Acoustic model	**Training data** (weights in loss function)	**SER%**
DNN_baseline [5]	CanAB	48.08
MT-TDNN-BLSTM	CanAB, K086, CUSENT (0.60, 0.30, 0.10)	38.64
MT-TDNN-BLSTM	CanAB, K086, CUSENT (0.65, 0.10, 0.25)	38.74
MT-TDNN-BLSTM	CanAB, K086, CUSENT (0.65, 0.25, 0.10)	**38.05**

Table 1 shows the overall SERs obtained by above ASR systems. We see that the MT-TDNN-BLSTM based systems significantly outperform the DNN_baseline system, achieving a SER reduction up to **10.03%**. The noticeable SER reduction shows the benefits of the large amount of training data as well as the effectiveness of proposed MT-TDNN-BLSTM structure and acoustic features. Besides, it is observed that treating the K086 as the secondary task (with submaximal weight value) performs better than that as the tertiary task. For the MT-TDNN-BLSTM system with the lowest SER, the SER per speaker varies greatly from 13.54% to 92.13% due to the highly diverse types and degree of language impairment. Across the 7 speech tasks, the ASR system shows similar SER performance, i.e., 36.78% to 39.54%. The outputs of this recognizer will be used for the following feature extraction procedure.

4. Feature extraction

4.1. Text features: Syllable-level embedding features derived from N-best hypotheses and confusion networks

We aim at text features that can be derived from the ASR output, and robust to the recognition errors as much as possible. In previous study [5], we provided a novel approach to extracting ASR-derived text features based on word embedding techniques. For a given story, a compact story-level vector representation was obtained by taking the average of all syllable vectors in accordance to the 1-best ASR output of this story. The syllable vectors were obtained from a continuous bag-of-words model [31] trained with syllable-level transcriptions of the training set of CanAB. Two text features were designed to quantify the degree of language impairment for each subject. **Inter-story** feature was able to capture the degree of confusion among 7 produced stories by counting the number of mis-clustered story vectors (up to 7, and normalized to the range of $[0, 1]$). If an aphasic subject produces mostly function words but few topic-specific content words, the value of inter-story feature would be high. **Intra-story** feature was defined as the cosine similarity between the story vector of an impaired speaker and the mean story vectors (with the same topic) of unimpaired ones, measuring the discrepancy between impaired and unimpaired content.

In present study, we propose to incorporate rich representation of ASR output (i.e. N-best hypotheses and confusion networks) into the computation of story vectors instead of using the straightforward 1-best ASR output. It is expected that the rich representation could provide a larger set of hypotheses to facilitate more robust story-level vector representations from erroneous ASR outputs.

4.1.1. Using N-best hypotheses for story vector computation

The output of an ASR system is typically the best-matching word sequence with the highest sentence-level posterior probability given an input speech utterance. In addition, the system can generate N-best candidates for each input utterance by ranking the combined acoustic model and language model (AM+LM) scores. In this study, after obtaining the 10-best hypotheses for each story, we propose to compute the new story-level vector representation by taking the following steps:
Step 1. For each story, 10 story vectors $\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_{10}$ are computed by averaging syllable vectors appeared in 10-best hypotheses respectively.
Step 2. Compute the AM+LM scores $C_1, C_2, \cdots, C_{10}$ for 10-best hypotheses using the function `nbest-to-linear` implemented in Kaldi. For each decoding hypothesis, its confidence score (weight) is computed by:

$$w_i = \frac{C_i}{C_1 + C_2 \cdots + C_{10}}, \quad i = 1, \ldots, 10; \tag{1}$$

Step 3. Compute the final weighted story vector as:

$$\boldsymbol{V}_{10-best} = w_1\boldsymbol{V}_1 + w_2\boldsymbol{V}_2 + \cdots + w_{10}\boldsymbol{V}_{10}. \tag{2}$$

4.1.2. Using confusion networks for story vector computation

Confusion networks (CNs) are direct linear graphical representations of most likely hypotheses in the lattice. Compared with the N-best lists, CNs contain more word candidates and sentence hypotheses. Figure 2 shows an example of CNs.

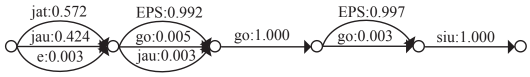

Figure 2: *Example of confusion networks for a speech segment.*

Each edge represents a syllable with its associated posterior probability. At each position, the sum of posterior probabilities of candidate syllables is equal to 1.0. "EPS" in the CNs represents a NULL hypothesis. In present study, we adopt the advanced method implemented in the Kaldi (by function `lattice-mbr-decode`) [32] to generate CNs. The story vectors combined 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 CN for a story and $N_1, \cdots, N_L$ denote the number of candidate syllables at each position. For the l^{th} position segment, the candidate syllables are $\omega_{1,l}, \cdots \omega_{N_L,l}$ with posterior probabilities $p_{1,l}, \cdots p_{N_L,l}$;
Step 2. For each story, the story vector is computed as the weighted average of all candidate syllable vectors appeared in the corresponding CN. The weight for each candidate syllable corresponds to the posterior probability generated from CN. The final story vector is computed using the following equation:

$$\boldsymbol{V}_{cn} = \frac{\sum_{l=1}^{L}\sum_{i=1}^{N_l} p_{i,l}\boldsymbol{V}(\omega_{i,l})}{L - L_{EPS}}, \tag{3}$$

where $\boldsymbol{V}(\omega_{i,l})$ indicates the syllable vector of $\omega_{i,l}$. 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.

After obtaining the story vectors derived from N-best hypotheses and CNs, the inter-story and intra-story features are computed following the methods in previous study [5]. We omit the details here due to the space limitation.

4.1.3. Text features of impaired speech: the effect of ASR

As shown in Table 2, we divide the 82 aphasic speakers into two groups with the SER below 50% and over 50%. For each group, we compare the average deviations of text features computed based on 1-best hypothesis, 10-best hypotheses and CNs from those based on manual transcriptions respectively. For the inter-story feature, a positive deviation means the number of mis-clustered story vectors tend to be over-estimated based on the ASR outputs. For the intra-story feature, a negative deviation indicates over-estimated discrepancy between impaired and unimpaired content. Both of them would lead to the over-estimation of impairment severity.

Table 2: *Deviations of text features computed from various ASR outputs from those from manual transcriptions in two groups.*

<table>
<tr><td colspan="2"><b>Acoustic Model</b></td><td colspan="6"><b>DNN_baseline</b> [5]</td></tr>
<tr><td colspan="2">SER</td><td colspan="3">SER $\leq 50\%$</td><td colspan="3">SER $> 50\%$</td></tr>
<tr><td colspan="2">No. of speakers</td><td colspan="3">49</td><td colspan="3">33</td></tr>
<tr><td colspan="2">Story vector derived from</td><td colspan="6">1-best</td></tr>
<tr><td rowspan="2">Deviation of feature values</td><td>inter-story</td><td colspan="3">0.020</td><td colspan="3">0.182</td></tr>
<tr><td>intra-story</td><td colspan="3">0.002</td><td colspan="3">−0.092</td></tr>
<tr><td colspan="2"><b>Acoustic Model</b></td><td colspan="6"><b>MT-TDNN-BLSTM</b></td></tr>
<tr><td colspan="2">SER</td><td colspan="3">SER $\leq 50\%$</td><td colspan="3">SER $> 50\%$</td></tr>
<tr><td colspan="2">No. of speakers</td><td colspan="3"><b>56</b></td><td colspan="3"><b>26</b></td></tr>
<tr><td colspan="2">Story vector derived from</td><td>1-best</td><td>10-best</td><td>CNs</td><td>1-best</td><td>10-best</td><td>CNs</td></tr>
<tr><td rowspan="2">Deviation of feature values</td><td>Inter-story</td><td><b>0.003</b></td><td><b>0.003</b></td><td><b>0.000</b></td><td><b>0.104</b></td><td><b>0.099</b></td><td><b>0.094</b></td></tr>
<tr><td>Intra-story</td><td>0.002</td><td>0.005</td><td>0.010</td><td><b>−0.078</b></td><td><b>−0.072</b></td><td><b>−0.062</b></td></tr>
</table>

As a consequence of the improvement of ASR performance with MT-TDNN-BLSTM, the number of impaired speakers in low-SER group increases from 49 to 56. For the text features computed from 1-best ASR output, it is shown that almost all deviations for two groups decrease significantly using the MT-TDNN-BLSTM system compared with DNN_baseline. This may benefit from more reliable ASR outputs produced by the MT-TDNN-BLSTM system. For the group of high-SER subjects, the text features derived from both 10-best hypotheses and CNs deviate less than those from 1-best hypothesis. Also, the CNs perform better than the 10-best hypotheses. For subjects with low SER, the average deviation of inter-story feature derived from CNs is smaller than that from 1-best hypothesis, but the deviation of intra-story feature is slightly higher. Overall speaking, using the MT-TDNN-BLSTM based recognizer and weighted story vectors computed from rich representation of ASR output can improve the robustness of text features to the ASR errors. This would alleviate the over-estimation of impairment severity, especially for impaired subjects with poor ASR performance.

4.2. Acoustic features: Supra-segmental Duration

Supra-segmental duration features derived from the ASR-generated time alignment were shown to provide additional benefit to the assessment system [5]. Based on previous studies on acoustical analysis of aphasia [4, 33], we defined 13 duration features that were related to fluency, speaking rate, etc. It is believed that the time-alignment produced by MT-TDNN-BLSTM system should be more accurate than that produced by the DNN_baseline system. The correlations between duration features generated from MT-TDNN-BLSTM system and the AQ values indeed increase significantly than before. In order to select the most effective features and reduce the feature dimension for the subsequent regression process, we jointly consider the ranking order of 13 candidate features suggested by the LASSO regression [34, 35] and the correlation between each feature and the AQ value. Finally, the following 4 duration features are selected:

Average duration of speech segments – Average duration of all speech segments, computed over all stories from the subject. Each speech segment is the part between two successive silence segments (> 0.5s);

Average syllable count per speech segment – The number of syllables per speech segment is computed and averaged over all stories from the subject;

Ratio of silence segment count to syllable count – The ratio of the number of silence segments to the number of syllables in all stories from the subject;

Proportion of fillers per speech segment – The proportion of the number of fillers per speech segment in all stories from the subject.

5. Automatic prediction of AQ

Automatic prediction of AQ is framed by performing regression on the feature vectors. Two regression models are built with the approaches of linear regression (LR) and random forest (RF) respectively. The leave-one-out cross validation strategy is adopted. In each fold of validation, the 6-dimensional feature vector from one of the subjects is reserved as test data and the remaining feature vectors are used for training. The feature vector consists of 2-dimensional text feature vector and 4-dimensional acoustic feature vector. Text features derived from 1-best hypothesis, 10-best hypotheses and CNs of ASR systems are evaluated separately. Table 3 shows the Spearman correlations between the predicted AQ (AQ_p) and the reference AQ (AQ_r) values obtained from the two regression models.

Table 3: *Correlations of predicted AQ with reference AQ values.*

Text features derived from	**LR**	**RF**
1-best of DNN_baseline [5]	0.816	0.839
1-best of MT-TDNN-BLSTM	**0.819**	0.839
10-best of MT-TDNN-BLSTM	**0.825**	**0.841**
CNs of MT-TDNN-BLSTM	**0.827**	**0.842**

Considering the case of using 1-best ASR output to extract text features, under the LR model, it is found that the MT-TDNN-BLSTM based ASR system performs better on AQ prediction than the DNN_baseline system. It is because the automated transcriptions and time alignment information from MT-TDNN-BLSTM system are more reliable. The results also show that the proposed robust text features derived from 10-best hypotheses and CNs can provide additional benefit to the AQ prediction. The largest gain is obtained by using the CNs for the computation of text features. The best prediction result shows a correlation of 0.842 based on the RF regression model. With this model, 29.3% (24/82) of the aphasic subjects have the prediction errors $|AQ_p - AQ_r| \leq 3.0$. There are 53.7% (44/82) and 75.6% (62/82) subjects have the prediction errors smaller than 6.0 and 10.0 respectively.

6. Conclusions

This paper presented an investigation on automatic speech assessment for PWA using the MT-TDNN-BLSTM based ASR system. Until this study, we have made good progress in improving the ASR performance on impaired speech and the robustness of text features. The performance of our assessment system is enhanced accordingly. In the future, we may also take the syntactic impairment of aphasic speakers into consideration in the assessment system.

7. 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] K. C. Fraser, F. Rudzicz, and E. Rochon, “Using text and acoustic features to diagnose progressive aphasia and its subtypes.” in *Proc. INTERSPEECH*, 2013, pp. 2177–2181.

[3] 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.

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

[5] 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.

[6] Y. Qin, T. Lee, A. P. H. Kong, and S. P. Law, “Towards automatic assessment of aphasia speech using automatic speech recognition techniques,” in *Proc. ISCSLP*, 2016, pp. 1–4.

[7] H. Christensen, M. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, “Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech.” in *Proc. INTERSPEECH*, 2013, pp. 3642–3645.

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

[9] R. Caruana, “Multitask learning,” in *Learning to learn*. Springer, 1998, pp. 95–133.

[10] S. Feng and T. Lee, “Improving cross-lingual knowledge transferability using multilingual TDNN-BLSTM with language-dependent pre-final layer,” in *Accepted by INTERSPEECH*, 2018.

[11] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. INTERSPEECH*, 2015, pp. 3214–3218.

[12] H. Sak, A. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” in *Proc. INTERSPEECH*, 2014, pp. 338–342.

[13] A. Graves, N. Jaitly, and A.-r. Mohamed, “Hybrid speech recognition with deep bidirectional LSTM,” in *Proc. ASRU*, 2013, pp. 273–278.

[14] P. Smit, S. Gangireddy, S. Enarvi, S. Virpioja, and M. Kurimo, “Aalto system for the 2017 Arabic multi-genre broadcast challenge,” in *Proc. ASRU*, 2017, pp. 353–359.

[15] A. Stolcke, Y. Konig, and M. Weintraub, “Explicit word error minimization in N-best list rescoring,” in *Proc. EUROSPEECH*, 1997, pp. 163–166.

[16] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus among words: Lattice-based word error minimization,” in *Proc. EUROSPEECH*, 1999, pp. 495–498.

[17] K. C. Sim, W. J. Byrne, M. J. Gales, H. Sahbi, and P. C. Woodland, “Consensus network decoding for statistical machine translation system combination,” in *Proc. ICASSP*, 2007, pp. IV–105.

[18] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur, “Beyond ASR 1-best: Using word confusion networks in spoken language understanding,” *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.

[19] A. P.-H. Kong and S.-P. Law, “The Cantonese Aphasiabank.” [Online]. Available: http://www.speech.hku.hk/caphbank/search/

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

[21] 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.

[22] T. Lee, W. K. Lo, P. Ching, and H. Meng, “Spoken language resources for Cantonese speech processing,” *Speech Communication*, vol. 36, no. 3-4, pp. 327–342, 2002.

[23] SpeechOcean, “King-ASR-086.” [Online]. Available: http://kingline.speechocean.com/exchange.php?id=1531&act=view

[24] G. Pironkov, S. Dupont, and T. Dutoit, “Multi-task learning for speech recognition: an overview,” in *Proc. ESANN*, vol. 192, 2016.

[25] 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.

[26] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. ICML*, 2015, pp. 448–456.

[27] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[28] D. Povey, X. Zhang, and S. Khudanpur, “Parallel training of deep neural networks with natural gradient and parameter averaging,” *arXiv preprint*, 2014.

[29] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *Proc. ICASSP*, 2014, pp. 2494–2498.

[30] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, “An exploration of dropout with LSTMs,” in *Proc. INTERSPEECH*, 2017, pp. 1586–1590.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint:1301.3781*, 2013.

[32] H. Xu, D. Povey, L. Mangu, and J. Zhu, “Minimum bayes risk decoding and system combination based on a recursion for edit distance,” *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[33] T. Lee, A. Kong, V. Chan, and H. Wang, “Analysis of auto-aligned and auto-segmented oral discourse by speakers with aphasia: A preliminary study on the acoustic parameter of duration.” *Procedia, social and behavioral sciences*, vol. 94, pp. 71–72, 2013.

[34] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[35] A. Loukina, K. Zechner, L. Chen, and M. Heilman, “Feature selection for automated speech scoring.” in *BEA@ NAACL-HLT*, 2015, pp. 12–19.