DEVELOPMENT OF THE CUHK ELDERLY SPEECH RECOGNITION SYSTEM FOR NEUROCOGNITIVE DISORDER DETECTION USING THE DEMENTIABANK CORPUS

Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, Helen Meng

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong, China

{zye,skhu,jcli,mzgeng,jwyu,jhxu,byxue,ssliu,xyliu,hmmeng}@se.cuhk.edu.hk xr.xie@link.cuhk.edu.hk

ABSTRACT

Early diagnosis of Neurocognitive Disorder (NCD) is crucial in facilitating preventive care and timely treatment to delay further progression. This paper presents the development of a state-of-the-art automatic speech recognition (ASR) system built on the DementiaBank Pitt corpus for automatic NCD detection. Speed perturbation based audio data augmentation expanded the limited elderly speech data by four times. Large quantities of out-of-domain, non-aged adult speech were exploited by cross-domain adapting a 1000-hour LibriSpeech corpus trained LF-MMI factored TDNN system to DementiaBank. The variability among elderly speakers was modeled using i-Vector and learning hidden unit contributions (LHUC) based speaker adaptive training. Robust Bayesian estimation of TDNN systems and LHUC transforms were used in both cross-domain and speaker adaptation. A Transformer language model was also built to improve the final system performance. A word error rate (WER) reduction of 11.72% absolute (26.11% relative) was obtained over the baseline i-Vector adapted LF-MMI TDNN system on the evaluation data of 48 elderly speakers. The best NCD detection accuracy of 88%, comparable to that using the ground truth speech transcripts, was obtained using the textual features extracted from the final ASR system outputs.

Index Terms— Automatic Speech Recognition, Elderly Speech, Neurocognitive Disorder Detection, Dementia

1. INTRODUCTION

Ageing presents enormous challenges to health care worldwide. Neurocognitive disorders (NCDs), such as Alzheimer's disease (AD), are often found among older adults [1]. Mild Cognitive Impairment (MCI) is an insidious and preclinical phase of AD and other forms of NCD. It is followed by a gradual progression of neurocognitive decline leading to an irreversible deterioration in memory, communication, orientation, and learning. Early diagnosis of NCD is crucial in facilitating preventive care and timely treatment to delay further progression and the occurrence of new symptoms [2]. NCDs like MCI often manifest themselves in speech and language impairments including weakened neuro-motor control in speech production and imprecise articulation, diminishing ability in using and comprehension of language, reduced vocabulary coverage, grammatical structure, as well as increasing difficulty in listening, reading and writing. Compared with other screening techniques based on brain scans or blood tests, speech and language based NCD diagnosis provides a non-intrusive alternative.

Traditionally such early screening tests are conducted manually by clinical professionals via neuropsychological tests. Large scale manual NCD screening among the elderly is difficult due to the relative shortage of clinical professionals and inter-rater variability in the assessment. One solution to this problem is to use fully automated machine learning based spoken language analytic approaches to improve the scalability of NCD screening for large population groups [3, 4]. In this process, a rich set of acoustic, articulatory, phonetic, prosodic, lexical, syntactic, and semantic level features encoding vital cues on speech and language deficiencies need to be extracted for NCD diagnosis [5, 6, 7, 8, 9, 10, 11]. As manually transcribing the conversation with the elderly subject and extracting these features is impractical on a large scale, automatic speech recognition (ASR) technologies tailed designed for elderly speech can be used.

Older adult speech exhibits a wide spectrum of new challenges for ASR system development. First, increased voice perturbations, articulatory imprecision, reduced speaking rates, increasing dysfluencies and decreasing intensities create a large mismatch between non-aged adult and elderly speech. The progression of NCDs such as Alzheimer's disease and other forms of dementia further aggregate speech and language deficiencies and diversities among the elderly people. State-of-the-art off-shelf commercial speech recognition systems designed for non-aged adult speech often produce high recognition error rates when directly used on elderly speech [9, 10, 12, 13, 14]. Second, it is also difficult to collect large amounts of speech recordings from face-to-face neuropsychological tests. For data-intensive deep learning technologies that are the staple modeling choice in current ASR systems [15, 16], large quantities of such well-matched, in-domain speech data are essential for system development. For these reasons, there has been limited research conducted on speech recognition system development for NCD screening. In order to address the data scarcity issue, in-house collected data was previously exploited in system development [4, 17]. However, compared with the rapid progress of speech recognition performance on non-aged adult speech, there is a notable lack of systems purposefully developed for elderly speech data targeting NCD diagnosis [4, 17, 18, 19, 20]. In particular, very limited speech recognition research [4, 19] has been conducted on the DementiaBank data [21], the largest publicly available speech corpus for NCD research.

In order to address these issues, this paper presents an initial attempt at the Chinese University of Hong Kong to design ASR system using a 33-hour DementiaBank Pitt corpus. Speech segmentation extracted from the original transcripts was first refined by removing excessive silence from each utterance. Speed perturbation based data augmentation methods [22] were used to expand the limited elderly training data by a factor of 4 times. State-of-the-art hybrid DNN-HMM systems featuring lattice-free maximum mutual information (LF-MMI) criterion [23] based sequence discrim-

Table 1: Statistics of the training, development, evaluation sets of the Pitt corpus used in this paper for ASR system development in terms of the number of elderly participants (PAR) and the number of hours of the speech recorded for both the participant and the investigator (INV), before (Column 3-5) and after (Column 6-8) audio re-segmentation was performed

	#PAR	Before Audio Re-segmentation			After Audio Re-segmentation		
		PAR	INV	Total	PAR	INV	Total
Train	244	17.65h	9.51h	27.16h	9.71h	6.03h	15.74h
Dev.	43	2.96h	1.79h	4.75h	1.40h	1.12h	2.52h
Eval.	48	0.88h	0.19h	1.07h	0.53h	0.09h	0.62h

inatively trained time delay neural network (TDNN) [15] acoustic models were developed. In order to further exploit out-of-domain, non-aged adult speech available in large quantities, a 1000-hour LibriSpeech corpus trained LF-MMI TDNN system is rapidly cross-domain adapted to the in-domain DementiaBank data. The variability among elderly speakers in both the original and augmented data was modeled using learning hidden unit contributions (LHUC) [24, 25] based speaker adaptive training. In order to account for the model uncertainty resulting from insufficient elderly speech data in the cross-domain and speaker adaptation stages, Bayesian estimation of the LF-MMI TDNN system parameters [16] and the speaker dependent LHUC transforms [26] were further exploited. Transformer language model [27, 28] was also used to improve the final system performance. An word error rate (WER) reduction of 11.72% absolute (26.11% relative) was obtained on the evaluation data consisting of 48 elderly speakers. The resulting systems' recognition outputs were then used to extract textual features for downstream NCD detection task [29]. An analysis of the correlation between speech recognition accuracy and NCD detection performance is presented.

The main contributions of this paper are summarized here. To the best of our knowledge, this is the first work to design state-of-the-art deep learning based ASR systems on the DementiaBank corpus for automatic NCD screening. In contrast, the previous research used either off-shelf commercial speech recognition systems [9, 10, 14], or more traditional GMM-HMM models in system development [4, 17, 18, 19] using a mix of publicly available and in-house datasets.

The rest of this paper is organized as follows. Section 2 introduces the data and the baseline system used. Section 3 describes the detailed development of the recognition system. Section 4 shows the NCD detection system performance using ASR outputs. Finally, the conclusions are drawn and future works are discussed in Section 5.

2. TASK DESCRIPTION

This section describes the audio and text data used in this paper and the baseline system structure.

Audio Data: In this paper, the Pitt corpus[1] [21] from DementiaBank was used in ASR system training. The Pitt corpus contains about 33-hour audios recorded over interviews between the 292 elderly participants and the clinical investigators. The word-level transcripts with approximate utterance-level segmentation are provided and all the elderly participants are labeled as either control, AD, or MCI, etc. The Pitt corpus was further split into the training, development and evaluation sets for building the ASR systems. The details regarding each of the three sets, in terms of the number of the participants and the total number of hours recorded for the participants and the investigators, are shown from Column 2-5 in Table 1. It should be noted that the evaluation set is exactly based on the same 48 speakers' Cookie section recordings as the ADReSS[2] [30] test set, while the development set contains the remaining recordings of the same speakers in other task sections if available.

Text Data: For language model, a mixture of text corpora was used, including the English transcripts (167k words) of DementiaBank (Pitt [21], Holland [31], Kempler [32], Lanzi [33]), the LDC Switchboard[3] and Fisher (LDC2004T19,LDC2005T19) telephone conversation transcripts (23.7m words) [34, 35], the New York Times Newswire Service (137.8m words) and Los Angeles Times/ Washington Post Newswire Service (254.6m words) portions from the LDC 5th edition Gigaword corpus (LDC2011T07) [36]. Two 4-gram language models with modified Kneser-Ney smoothing were constructed using the SRILM toolkit [37]. The first "small" 4-gram LM was built using the Pitt data only, while the other "large" 4-gram LM was constructed using the probability level linear interpolation over component models trained on each of the text sources mentioned above separately before being combined. These two language models were used in most of the experiments in Table 2. A 3.6k word recognition vocabulary covering all the words in the Pitt corpus with standard American phonetic pronunciation was used.

Baseline System: LF-MMI sequence trained hybrid TDNN models were built [15, 23]. Following the Kaldi recipe[4], a GMM-HMM system was first built with 1800 tied tri-phone states with 32 Gaussians each, using Maximum Likelihood Linear Transform (MLLT) [38, 39] on the Linear Discriminant Analysis (LDA) transformed 39-dim Perceptual Linear Prediction (PLP) coefficients, including differential parameters up to the second order. Speaker adaptive training (SAT) [40] system was used to generate the alignments and the finite-state transducer (FST) lattices for LF-MMI training. A 14-layer TDNN was then trained using one thread only on a single NVIDIA V100 GPU with 40-dim filterbank input features. Statistical significance test was conducted at level $\alpha = 0.5$ based on matched pairs sentence-segment word error (MAPSSWE) for recognition performance analysis.

3. SPEECH RECOGNITION SYSTEM DEVELOPMENT

This section presents the performance of the baseline TDNN systems before introducing a series of techniques to further improve the recognition accuracy. The overall architecture is shown in Figure 1.

3.1. Baseline System Performance

The performances of the TDNN systems using the original segmentation of the Pitt corpus with or without i-Vector [41] and optionally using small or large language models described in Section 2 are shown in Table 2. It was found that using i-Vector provided marginal improvement. For example, from Sys. 3 to Sys. 4, the word error rate (WER) was reduced by 0.19% absolute only. The large language model further improved the system performance. For example, the WER was reduced by 0.69% absolute (1.07% absolute for participants in the evaluation set) from Sys. 2 to Sys. 4. The best result of the baseline systems was obtained by incorporating i-Vector adaptation and the large language model (Sys. 4).

[1] https://dementia.talkbank.org/access/English/Pitt.html

[2] http://www.homepages.ed.ac.uk/sluzfil/ADReSS/

[3] http://www.isip.piconepress.com/projects/switchboard/releases/switchboard_word_alignments.tar.gz

[4] Kaldi: egs/swbd/s5c/local/chain/tuning/run_tdnn_7q.sh

Table 2: Word error rate (WER%) of DementiaBank Pitt development and evaluation sets obtained using the baseline systems with or without i-Vector and optionally using the small or large language models (Sys. 1-4); the performance of the TDNN systems improved through different stages: A. audio re-segmentation (Sys. 5); B. speed perturbation based data augmentation (Sys. 6); C. domain adaptation (Sys. 7); D. speaker adaptation (Sys. 8-9); E. transformer language model re-scoring (Sys. 10), with "PAR" for participant and "INV" for investigator. The "small 4-gram" was trained with the Pitt data only while the "large 4-gram" incorporated other corpora. † denotes statistical significant difference in result is obtained compared with the baseline system (Sys. 4)

Sys.	I-Vector	Audio Re-segment	Speed perturb	Bayesian TDNN Adaptation		Language Model	Dev.		Eval.		All
				Domain	Speaker		PAR	INV	PAR	INV	
1	×	×	×	×	×	small 4-gram	53.48	22.65	43.54	29.06	38.53
2	√					small 4-gram	52.93	23.16	45.96	27.62	38.87
3	×					large 4-gram	52.87	23.15	43.00	28.18	38.37
4	√					large 4-gram	51.70	23.13	44.89	26.85	38.18
5	√	√	×	×	×	large 4-gram	51.51	21.57	39.01	20.64	36.31†
6	√	√	√	×	×	large 4-gram	46.76	19.97	37.01	18.20	33.37†
7	√	√	√	√	×	large 4-gram	45.56	19.19	35.31	19.31	32.33†
8				×	BLHUC-SAT		42.95	18.24	34.12	17.87	30.67†
9				√	BLHUC-SAT		43.74	18.06	33.82	**16.65**	30.82†
10	√	√	√	√	BLHUC-SAT	large 4-gram + Transformer	**42.12**	**17.61**	**33.17**	17.20	**29.90**†

3.2. Audio Re-segmentation

In order to improve the original audio segmentation provided by the Pitt corpus, a GMM-HMM system with 2k tied tri-phone states and 32 Gaussians per state was trained with HTK toolkit [42] to force align the training, development and evaluation sets. During the re-segmentation stage, excessive silences longer than 200ms at the start or the end of each utterance were removed. Long utterances containing sentence internal pauses longer than 1 second were further split into multiple shorter utterances. Compared with the quantity of data using the original transcripts, the re-segmentation stage reduced the training set from 27.16 hours to 15.74 hours as shown in Table 1. Similar reduction ratios of duration were also observed in the development and evaluation sets. Using the refined audio segmentation, the resulting TDNN system (Sys. 5) outperformed the baseline system (Sys. 4) with a statistically significant WER reduction of 1.87% absolute (5.88% absolute for participants in the evaluation set).

3.3. Data Augmentation

In order to expand the limited training data of only 15.74 hours after re-segmentation, following the previous research on data augmentation for normal speech [43] and disordered speech [22], speed perturbation based data augmentation was subsequently performed. First, speaker independent speed perturbation with a fixed perturbation factor set $\{0.9, 1.0, 1.1\}$ was used to expand the participants' speech data by a factor of 3 to about 29 hours. Second, the investigators' speech were further speed perturbed with a different factor set $\{0.84, 0.95, 1.0, 1.08, 1.27\}$. These were initially computed at the speaker level for each elderly participant relative to the average investigator's speaking rate using phonetic alignment analysis [44], before K-means clustering was used to group the speaker-level perturbation factors into the above set of five values. This transformed the 6-hour investigators' speech into approximately 30 hours of the elderly participant like speech. Combining the augmented data from the two stages above, the total amount of the training data was increased from 15.75 hours by a factor of 4 to roughly 59 hours. Using the augmented training data, the resulting system (Sys. 6) outperformed the baseline system (Sys. 5) by 2.94% absolute (2% absolute for participants in the evaluation set) in terms of WER reduction.

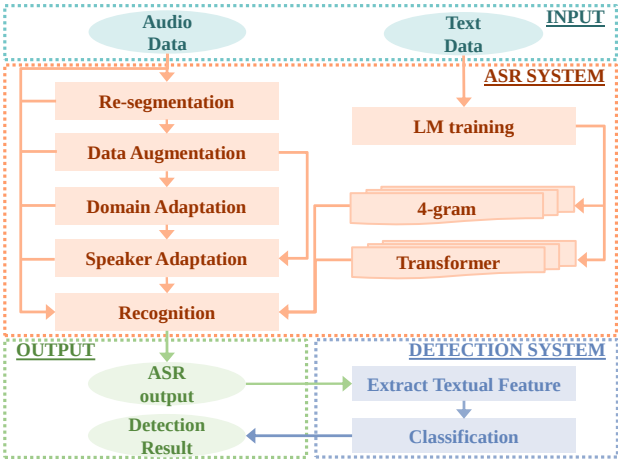

Fig. 1: The overall speech recognition system (Sec. 3) and NCD detection system (Sec. 4) architecture considered in this paper

3.4. Cross-domain Adaptation

Bayesian learning provides a mathematically well-formulated framework to account for model uncertainty in a wide range of deep learning systems [16, 45] and has been successfully applied to improve the generalization performance of the LF-MMI sequence discriminatively trained TDNN acoustic models [16]. In order to exploit large quantities of out-of-domain, non-aged adult speech, we further explored the use of a Bayesian TDNN cross-domain adaptation approach. A 1000-hour LibriSpeech corpus trained LF-MMI TDNN system with Kaldi recipe[5] was rapidly cross-domain adapted to the 59-hour in-domain Pitt data after speed perturbation. During cross-domain adaptation, the first TDNN hidden layer, where the largest data diversity was expected compared with the higher layers producing more invariant features, was Bayesian adapted using a Gaussian parameter prior distribution centered around the parameters of the LibriSpeech corpus trained LF-MMI TDNN system fine-tuned to the Pitt data. The resulting Bayesian cross-domain adapted TDNN system (Sys. 7) outperformed the system trained on the Pitt data only (Sys. 6) by 1.04% absolute (1.7% absolute for participants in the evaluation set) in WER reduction.

[5] Kaldi: egs/librispeech/s5/{run.sh, local/chain/run tdnn.sh}

3.5. Speaker Adaptation

Individuals experiencing NCD at different stages of progression exhibit highly diverse voice characteristics. To this end, speaker adaptation techniques play a central role in reducing the mismatch between ASR systems and target elderly users. In order to robustly learn speaker-dependent adaptation parameters, DNN model based adaptation techniques, like learning hidden unit contributions (LHUC) [24, 25], often require a significant amount of speaker level enrollment data. In order to account for the model uncertainty resulting from the limited speaker-level adaptation data for each participant, Bayesian LHUC speaker adaptive training (SAT) [26] was further applied to model the large variability among elderly participants in both the original and the augmented Pitt training data. As shown in Table 2, the resulting Bayesian LHUC speaker adapted LF-MMI TDNN system (Sys. 8) outperformed the comparable speaker independent TDNN system (Sys. 6) by 2.7% absolute (2.89% absolute for participants in the evaluation set) in WER reduction. Further WER reductions of 0.3% absolute for participants and 1.22% absolute for investigators in the evaluation set were obtained by performing both domain and speaker adaptation (Sys. 9)[6].

3.6. Transformer Language Model

In order to further improve the generalization of the baseline 4-gram LMs, a Transformer language model [27, 28, 46] consisting of six stacked multiple self-attention layers followed by feedforward layers with residual connection and layer normalization inserted between them, as well as additional positional encoding layers, was trained on the combined 2.4m words of the DementiaBank Pitt, Switchboard and Fisher transcripts. In order to reduce the domain mismatch between the three text sources, the resulting Transformer was Bayesian adapted [47] to the Pitt transcripts only while serving as the Prior model. It was then linearly interpolated using equal weights with the large 4-gram LM to rescore the n-best lists produced by the domain and speaker adapted system (Sys. 9) in Table 2. Further absolute WER reduction of 0.92% absolute (0.65% absolute for participants in the evaluation set) was obtained over the 4-gram LM.

4. NCD DETECTION PERFORMANCE

In this section, the textual features separately extracted from the DementiaBank Pitt evaluation set recognition outputs produced by the baseline ASR system (Sys. 4) and the final recognition system (Sys. 10) in Table 2 respectively were fed into a Support Vector Machine (SVM) based NCD detection system, as illustrated in Figure 1. This detection system was maximum-margin trained on the ADReSS training set[7], a subset of the Cookie session transcripts of the Pitt corpus of 108 recordings [30]. Textual features based on either a) 1035-dim term frequency-inverse document frequency (TD-IDF) features [48], or b) 768-dim vector embeddings produced by a BERT model pre-trained on the English BookCorpus [46], were used with linear kernel, preprocessed with standard scaling and then Principle Component Analysis (PCA). More details of the detection system could be found in [29].

The results in Table 3 suggest the NCD detection performance based on the final ASR system was comparable to that using the ground truth speech transcripts with both textual features. With BERT based features, which may capture additional long-range contextual information such as syntactic structure complexity compared to the TF-IDF features encoding only word-level frequency information, the reduction of WER from 44.89% (Sys. 4) to 33.17% (Sys. 10) led to the improvements in NCD detection accuracy from 79% to 88%, F1 score from 0.82 to 0.88 and area under curve (AUC) from 0.87 to 0.92.

Table 3: ASR WER% and NCD detection performance in terms of accuracy, precision, recall F1 score and area under curve (AUC) obtained using the ground truth transcripts, the baseline or the best ASR outputs (Sys. 4 & 10 in Table 2) for participants of the evaluation set

Sys.	Feature	WER	Acc.	Pre.	Rec.	F1	AUC
Manual	TF-IDF	N/A	0.71	0.73	0.67	0.70	0.83
4		44.89	0.69	0.74	0.58	0.65	0.85
10		33.17	0.69	0.74	0.58	0.65	0.82
Manual	BERT	N/A	**0.88**	**0.91**	0.83	0.87	0.89
4		44.89	0.79	0.72	**0.96**	0.82	0.87
10		33.17	**0.88**	0.82	**0.96**	**0.88**	**0.92**

5. CONCLUSION

The development of a state-of-the-art ASR system constructed using the DementiaBank Pitt corpus was presented in this paper. A series of techniques featuring segmentation refinement, audio augmentation, Bayesian cross-domain and speaker adaptation as well as Transformer language models were employed to improve the recognition performance of elderly speech. An overall significant WER reduction of 11.72% absolute (26.11% relative) was obtained over the baseline i-Vector adapted LF-MMI TDNN system on the Pitt evaluation set consisting of 48 elderly speakers. The NCD detection performance using the textual features extracted from our ASR system outputs was also found comparable to that using the ground truth speech transcripts. Further analysis of the correlation between speech recognition accuracy and NCD detection performance was also presented. Tighter integration between the recognition and NCD detection components, fusion with paralinguistic features and further perturbation of the elderly speech with simulated noises and reverberation will be investigated in future research.

6. ACKNOWLEDGMENT

This research is supported by Hong Kong RGC GRF grant No. 14200218, 14200220, TRS T45-407/19N, Innovation & Technology Fund grant No. ITS/254/19, and SHIAE grant No. MMT-p1-19.

7. REFERENCES

[1] Alzheimer's Assoc., "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[2] C.P. Ferri, M. Prince, C. Brayne, et al., "Global prevalence of dementia: a delphi consensus study," *The lancet*, vol. 366, no. 9503, pp. 2112–2117, 2005.

[3] Y. Pan, B. Mirheidari, M. Reuber, et al., "Automatic hierarchical attention neural network for detecting ad.," in *INTERSPEECH*, 2019, pp. 4105–4109.

[4] B. Mirheidari, D. Blackburn, T. Walker, et al., "Dementia detection using automatic analysis of conversations," *CSPL*, vol. 53, pp. 65–79, 2019.

[5] K.C. Fraser, J.A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *J. Alzheimer's Dis.*, vol. 49, no. 2, pp. 407–422, 2016.

[6] Some degradation on the participant portion of the Dev set may be caused by the annotation errors found in the corresponding manual reference transcripts.

[7] The elderly participants labeled as memory, vascular, possible AD, probable AD are treated as NCD positive while control as NCD negative.

[6] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german.," in *INTERSPEECH*, 2016, pp. 1938–1942.

[7] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for alzheimer's disease," in *SLPAT*, 2016, pp. 32–36.

[8] B. Mirheidari, D. Blackburn, T. Walker, et al., "Detecting signs of dementia using word vector representations.," in *INTERSPEECH*, 2018, pp. 1893–1897.

[9] A. König, N. Linz, et al., "Fully automatic speech-based analysis of the semantic verbal fluency task," *Dementia and geriatric cognitive disorders*, vol. 45, no. 3-4, pp. 198–209, 2018.

[10] A. Pompili, A. Abad, et al., "Pragmatic aspects of discourse production for the automatic identification of alzheimer's disease," *IEEE JSTSP*, vol. 14, no. 2, pp. 261–271, 2020.

[11] R. Chakraborty, M. Pandharipande, et al., "Identification of dementia using audio biomarkers," *arXiv:2002.12788*, 2020.

[12] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of asr performance on ageing voices," 2008.

[13] D. Hakkani-Tür, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," in *INTERSPEECH*, 2010, pp. 258–261.

[14] K.C. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *SLPAT*, 2013, pp. 47–54.

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

[16] S. Hu, X. Xie, S. Liu, et al., "Lf-mmi training of bayesian and gaussian process time delay neural networks for speech recognition.," in *INTERSPEECH*, 2019, pp. 2793–2797.

[17] F. Rudzicz, R. Wang, M. Begum, and A. Mihailidis, "Speech recognition in alzheimer's disease with personal assistive robots," in *SLPAT*, 2014, pp. 20–28.

[18] M. Lehr, E. Prud'hommeaux, I. Shafran, et al., "Fully automated neuropsychological assessment for detecting mild cognitive impairment," in *INTERSPEECH*, 2012, pp. 1039–1042.

[19] L. Zhou, K.C. Fraser, and F. Rudzicz, "Speech recognition in alzheimer's disease and in its assessment.," in *INTERSPEECH*, 2016, pp. 1948–1952.

[20] L. Tóth, I. Hoffmann, G. Gosztolya, et al., "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[21] J.T. Becker, F. Boiler, O.L. Lopez, et al., "The natural history of alzheier's disease: description of study cohort and accuracy of diagnosis," *Arch. Neurol.*, vol. 51, no. 6, pp. 585–594, 1994.

[22] M. Geng, X. Xie, S. Liu, et al., "Investigation of data augmentation techniques for disordered speech recognition," in *INTERSPEECH*, 2020.

[23] D. Povey, V. Peddinti, D. Galvez, et al., "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *INTERSPEECH*, 2016, pp. 2751–2755.

[24] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE TASLP*, vol. 24, no. 8, pp. 1450–1463, 2016.

[25] P. Swietojanski and S. Renais, "Sat-lhuc: Speaker adaptive training for learning hidden unit contributions," in *ICASSP*. IEEE, 2016, pp. 5010–5014.

[26] X. Xie, X. Liu, T. Lee, S. Hu, and L. Wang, "Blhuc: Bayesian learning of hidden unit contributions for deep neural network speaker adaptation," in *ICASSP*. IEEE, 2019, pp. 5711–5715.

[27] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[28] K. Irie, A. Zeyer, et al., "Language modeling with deep transformers," in *INTERSPEECH*, 2019, pp. 3905–3909.

[29] J. Li, J. Yu, et al., "A comparative study of acoustic and linguistic features classification for alzheimer's disease detection," in *ICASSP*. IEEE, 2021.

[30] S. Luz, F. Haider, et al., "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," 2020.

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

[32] D. Kempler et al., "Syntactic preservation in alzheimer's disease," *JSLHR*, vol. 30, no. 3, pp. 343–350, 1987.

[33] A. Lanzi, S.E. Wallace, and M. Bourgeois, "Group external memory aid treatment for mild cognitive impairment," *Aphasiology*, vol. 33, no. 3, pp. 320–336, 2019.

[34] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *ICASSP*. IEEE, 1992, vol. 1, pp. 517–520.

[35] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[36] R. Parker and D. Graff, "Jumbo kong, ke chen, and kazuaki maeda. 2011," *English Gigaword 5th Edition. LDC*.

[37] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *INTERSPEECH*, 2002, pp. 901–904.

[38] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *CSPL*, vol. 9, no. 2, pp. 171–185, 1995.

[39] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *CSPL*, vol. 12, no. 2, pp. 75–98, 1998.

[40] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *ICASSP*. IEEE, 1997, vol. 2, pp. 1043–1046.

[41] N. Dehak, P.J. Kenny, R. Dehak, et al., "Front-end factor analysis for speaker verification," *IEEE TASLP*, vol. 19, no. 4, pp. 788–798, 2010.

[42] S. Young, G. Evermann, M. Gales, et al., "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, pp. 12, 2002.

[43] T. Ko, V. Peddinti, et al., "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015, pp. 3586–3589.

[44] F. Xiong et al., "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP*. IEEE, 2019, pp. 5836–5840.

[45] S. Hu, M. Lam, X. Xie, et al., "Bayesian and gaussian process neural networks for large vocabulary continuous speech recognition," in *ICASSP*. IEEE, 2019, pp. 6555–6559.

[46] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.

[47] B. Xue, J. Yu, et al., "Bayesian transformer language models for speech recognition," in *ICASSP*. IEEE, 2021.

[48] J. Ramos, "Using tf-idf to determine word relevance in document queries," .