FEATURE SELECTION AND TEXT EMBEDDING FOR DETECTING DEMENTIA FROM SPONTANEOUS CANTONESE

Xiaoquan Ke[1], *Man-Wai Mak*[1], *Helen M. Meng*[2]

[1]The Hong Kong Polytechnic University, Hong Kong SAR
[2]The Chinese University of Hong Kong, Hong Kong SAR

ABSTRACT

Dementia is a severe cognitive impairment that affects the health of older adults and creates a burden on their families and caretakers. This paper analyzes diverse hand-crafted features extracted from spoken languages and selects the most discriminative ones for dementia detection. Recently, the performance of dementia detection has been significantly improved by utilizing Transformer-based models that automatically capture the structural and linguistic properties of spoken languages. We investigate using Transformer-based features and propose an end-to-end system for dementia detection. We also explore recent ASR and representation learning frameworks, such as Wav2vec 2.0 and Hubert, for transcribing the Marvel Cantonese corpus that contains recordings of older adults describing the rabbit story. We investigate using disfluency patterns (DP) in spontaneous speech to upgrade the input for Transformer-based classification. Results show that applying automatic transcriptions with DP to fine-tune Transformer-based models can improve dementia detection performance.

Index Terms— Dementia detection, Feature selection, ASR, Disfluency pattern, Transformer

1. INTRODUCTION

Dementia is the loss of cognitive functions (thinking, remembering, and reasoning) that seriously devastates the daily lives of the afflicted patients. The most common form of dementia is the Alzheimer's disease (AD), which may contribute to 60–70% of dementia cases. According to the World Alzheimer's Report,[1] more than 55 million people live with dementia worldwide, and there are nearly 10 million new cases every year. In 2019, the estimated global societal cost of dementia was $1.3 trillion, and these costs are expected to surpass $2.8 trillion by 2030. The disease has a huge impact on the quality of life of not only the patients but also their families and caretakers. Fortunately, with effective detection of early dementia, disease-modifying medications and interventions are possible [1].

1.1. Related Work

Recently, automatic detection of dementia through speech and language analyses has gathered attention in the research community. Some studies investigated different types of speech-based features for dementia detection. For example, some studies used acoustic information (e.g., speech/silence segments and voice quality [2]) from speech waveforms to discover potential dementia. More recently, Haider *et al.* [3] compared different types of paralinguistic features – including eGeMAPS [4], ComParE 2013 [5], Emobase [5], and MRCG [6] – for dementia detection. As the paralinguistic features are high-dimensional, Pearson's correlation (PeaCorr) tests were performed to reduce the feature dimensions.

In addition to speech-based features, transcription-based features have also been used for dementia detection. These features can be extracted from automatic or manual transcriptions, which capture the semantic, syntactic, and lexical aspects of the speaker's utterances. For example, Qiao *et al.* [7] combined disfluency and linguistic complexity features for AD detection. The linguistic complexity features (syntactic complexity, lexical richness, register-based n-gram frequency, and information-theoretic measures) were generated by analyzing the transcriptions using the Complexity Contour Generator (CoCoGen) [8].

1.2. Modeling Approach

The modeling approach presented in this paper builds on key insights from the above studies by combing hand-crafted features and text embeddings for dementia detection. The text embeddings are extracted from transformer-based models. For example, in [7], the BERT [9] model was fine-tuned to capture the language characteristics of AD patients. We build an end-to-end system containing two branches to thoroughly model the hand-crafted features and text embeddings, as shown in Figure 1. Branch 1 extracts hand-crafted features, followed by feature selection (FS) to select the discriminative features. Branch 2 is built on text embeddings. We obtain the final score for the whole speech recording by averaging the scores from the two branches. The proposed system is evaluated on a Cantonese corpus called Marvel.

[1] `https://www.who.int/news-room/fact-sheets/detail/dementia`

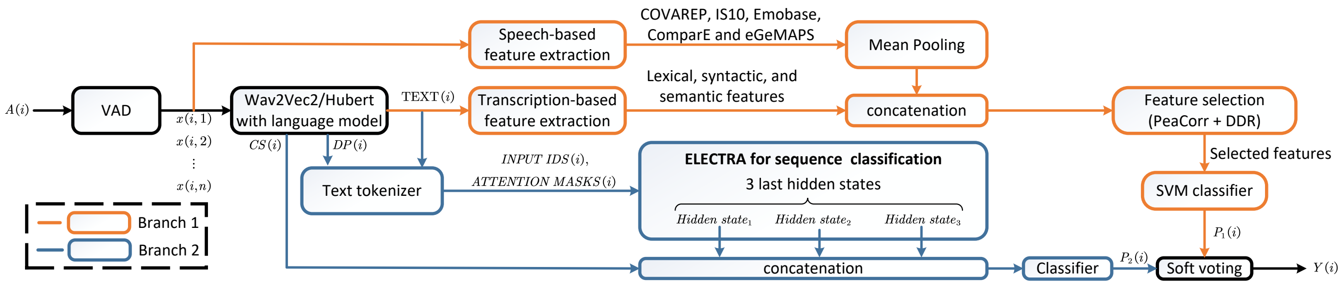
$A(i)$
VAD
$x(i,1)$
$x(i,2)$
⋮
$x(i,n)$
Wav2Vec2/Hubert with language model
TEXT (i)
$CS(i)$
$DP(i)$
Speech-based feature extraction
COVAREP, IS10, Emobase, Compar E and eGeMAPS
Mean Pooling
Transcription-based feature extraction
Lexical, syntactic, and semantic features
concatenation
Feature selection (PeaCorr + DDR)
Selected features
SVM classifier
$P_1(i)$
Text tokenizer
$INPUT\ IDS(i)$, $ATTENTION\ MASKS(i)$
ELECTRA for sequence classification
3 last hidden states
$Hidden\ state_1$
$Hidden\ state_2$
$Hidden\ state_3$
concatenation
Classifier
$P_2(i)$
Soft voting
$Y(i)$
Branch 1
Branch 2

Fig. 1. Our end-to-end system to detect dementia from spontaneous Cantonese. The recording $A(i)$ is segmented using voice activity detection (VAD) into n segments. Branch 1 is built on hand-crafted features, and Branch 2 is built on text embeddings. CS: Confidence score; DP: Disfluency pattern. The final label $Y(i)$ for the whole speech recording is obtained by averaging the scores from the two branches.

2. METHODS

2.1. Feature Selection on Hand-crafted Features

While various types of speech-based features and transcription-based features have been used for dementia detection, it is still unclear which features or their combinations are more effective. We built on key insights from the previous studies and utilized feature selection (FS) to find the most effective features for dementia detection. The transcription-based features (Figure 1, Branch 1) are described as follows. (1) *Lexical features.* With automated speech transcriptions, the Hanlp library was utilized to perform segmentation and part-of-speech (POS) tagging.[2] After that the following features were extracted: POS ratio, the ratio of pronoun to noun, and the ratio of noun to verb. We also measured the lexical richness by calculating the type-token-ratio. We counted the number of top-10 fillers in Cantonese and normalized it by the total number of word tokens in the transcriptions. (2) *Syntactic features.* We converted the transcriptions into simplified Chinese and measured the syntactic complexity in Chinese writing [10]. (3) *Semantic features.* Word specificity and ambiguity were computed based on tree depth and the number of senses in NLTK WordNet [11]. We then computed semantic similarity using the mean and minimum cosine distances between the one-hot embeddings of each pair of utterances [12].

The speech-based features (Figure 1, Branch 1) include COVAREP features [13] and four paralinguistic features sets, which are INTERSPEECH 2010 Paralinguistic Challenge Features (IS10) [14], Emobase [5], eGeMAPS [4], and ComParE [5].

We combined all the feature listed above and applied Dual-dropout ranking (DDR) [15] to rank and select features. We have applied DDR to select linguistic features for dementia detection in our previous research [15].

2.2. Text-Embedding Classification with Disfluency Pattern

We used the erroneous automatic transcriptions to build the text embeddings in Branch 2 of Figure 1. The erroneous transcriptions could impact the performance of dementia detection. To mitigate this problem, Pan *et al.* [16] used the confidence scores from an ASR system as a proxy measure for accuracy. They incorporated confidence scores into text embeddings, which provides the classifier with information about the transcription quality.

We also incorporated confidence scores into the text embeddings to mitigate the effect of erroneous transcriptions. We followed the structure in [16] and concatenated the last three hidden states of the ELECTRA[3] model with confidence scores as input to the classifier, as shown in Figure 1 (Branch 2). Additionally, we augmented the input of the ELECTRA model with multiple hypotheses generated by an ASR system. In addition to the best hypothesis, the ASR system with a language model can use different parameters to produce multiple hypotheses. With the range of language model's weight from 0.5 to 5 and the word score from 0 to 0.5, we produced 20 ASR hypotheses and confidence scores for each speech recording.

Disfluency–including silent pauses, filled pauses, repetitions, self-corrections, and discourse incoherence–is part of spontaneous speech. However, dementia patients manifest different patterns of disfluencies in spontaneous speech. For example, Yuan *et al.* [17] reported that AD patients have more pauses than healthy controls (HCs), especially the longer pauses. They coded short (under 0.5 sec), medium (0.5–2 sec), and long (over 2 sec) pauses using three punctuations ',' , '.', and '...' to expose the pauses in transcriptions. Their results demonstrate that using the transcriptions with pauses to fine-tune a BERT model can improve the performance of AD detection.

[2] https://github.com/hankcs/HanLP

[3] https://huggingface.co/toastynews/electra-hongkongese-base-discriminator

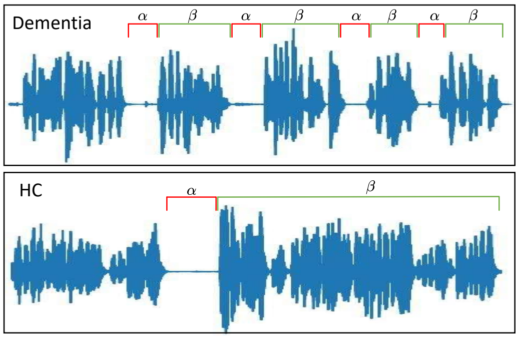

Fig. 2. The upper panel shows the disfluency pattern (DP) of a dementia patient, and the lower panel shows the DP of a healthy control. α and β refer to the time period in Stage 1 and Stage 2, respectively. It demonstrates that β/α of the dementia patient is quite small because the patient does not express enough content even after a long pause.

Inspired by [17], in addition to the pauses, we investigated the 'disfluency pattern' (DP) in spontaneous speech, as shown in Figure 2. We model the process of expressing spontaneous speech into two stages: Stage 1 (the pause) builds content in mind and Stage 2 expresses the content. The time period α in Stage 1 indicates how long a speaker takes to build the content. The time period β in Stage 2 indicates how much content a speaker has built in Stage 1. The β/α of the healthy control (HC) is large because the HC expresses lots of content after a period time of thinking. However, the β/α of the NCD patient is quite small because the patient does not express enough content even after a long pause, which indicates language impairment. The DP not only exposes the pauses but also exposes how much content a speaker has built in the period of pausing. The time alignment information from the ASR system was used for measuring α and β. If $\alpha > a$ and $\beta/\alpha < b$, we added a punctuation 'DP' to the transcriptions. We determined the best possible a and b using grid-search cross-validation (CV). The transcriptions with DP were used to fine-tune the ELECTRA model for dementia detection.

3. EXPERIMENTAL PROCEDURES

The Marvel dataset was collected by the CUHK for theme-based research (Ref.: T45-407/19-N). A series of cognitive tests including Montreal Cognitive Assessment (MoCA) tests and picture description tests were given to each participant for assessing the mild cognitive impairment (MCI) and dementia in older adults. According to the assessment results, 461 participants were divided into three groups: (1) 281 healthy older adults (HCs); (2) 144 older adults having minor neurocognitive disorders (minor NCD); and (3) 36 older adults suffering from major NCD.

For detecting dementia, we combined minor NCD and major NCD into one category called possible dementia. According to the age distribution and gender distribution, 120 participants (60 HCs and 60 possible dementia) were selected as the test data. A rabbit story picture description task was selected for the experiments. The performance metrics include accuracy (ACC), precision (PRE), recall (REC), and F_1 score with respect to the possible dementia category. The performance on the training data was obtained by 10-fold cross-validation (CV).

Because only a small subset of the Marvel dataset has manual transcriptions, automatic speech recognition (ASR) system was used for generating the transcriptions. Wav2vec 2.0 [18] (denoted as Wav2vec2 from now on) and Hubert [19] are both self-supervised training models that can be utilized for end-to-end ASR. Wav2vec2 and Hubert can learn powerful representations from a large amount of unlabeled speech data. By fine-tuning the models on a small amount of transcribed speech, Wav2vec2 and Hubert can achieve similar performance as traditional fully-supervised ASR systems. As there is no Cantonese pre-trained version of Wav2vec2 or Hubert, we adopted multilingual and Chinese pre-trained versions of Wav2vec2 and Hubert from the Transformer Python library, including *Wav2vec2-large-xlsr*,[4] *Wav2vec2-large-Chinese*,[5] and *Hubert-large-Chinese*.[6]

The Cantonese version of Common Voice Speech dataset [20] (common-voice-zh-HK) was used for fine-tuning. The PyCantonese library was utilized to convert the transcriptions to corresponding phone sequences.[7] The acoustic models were end-to-end fine-tuned on phone-level using connectionist temporal classification (CTC) loss. The fine-tuned acoustic models were tested on common-voice-zh-HK test data. The phone error rate (PER) for *Wav2vec2-large-xlsr*, *Wav2vec2-large-Chinese*, and *Hubert-large-Chinese* were 0.112, 0.183, and 0.107, respectively. Therefore, *Hubert-large-Chinese* was selected for transcribing the Marvel corpus. The outputs of the fine-tuned acoustic models were decoded using a beam search decoder with a 4-gram KenLM language model trained on common-voice-zh-HK.

4. RESULTS AND DISCUSSIONS

4.1. Performance of Feature Selection

We first evaluate the recognition performance of the full features *before* FS. We used a Gaussian SVM with $C = 1$ as the classifier to distinguish the possible dementia and the HCs, as shown in Table 1. Considering that the feature dimension is very high, filter methods were utilized to reduce the feature dimension before applying DDR to select features. On the

[4] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[5] https://huggingface.co/TencentGameMate/chinese-wav2vec2-large

[6] https://huggingface.co/TencentGameMate/chinese-hubert-large

[7] https://pycantonese.org/

Table 1. Classification performance of different feature types. The numbers in the brackets are the sizes of the feature sets.

Feature set	**10-fold CV on training data**		**Performance on test data**	
	ACC	F_1	ACC	F_1
Transcription-based (361)	0.678	0.577	0.667	0.638
COVAREP (518)	0.684	0.562	0.650	0.615
IS10 (1582)	0.704	0.599	0.692	0.670
Emobase (988)	0.705	0.605	0.675	0.645
eGeMAPS (88)	***0.720***	0.624	0.658	0.627
CompParE (6373)	0.707	0.588	0.667	0.641
All features (9910)	0.712	0.589	0.700	0.677
PeaCorr + DDR (18)	0.705	***0.628***	***0.708***	***0.699***

Table 2. Classification performance of different numbers of features selected by PeaCorr.

Feature dimension	**10-fold CV on training data**			
	ACC	PRE	REC	F_1
250	0.701	0.675	0.618	0.610
500	***0.708***	***0.685***	***0.622***	***0.623***
750	0.707	0.684	0.621	0.613
1,000	0.706	0.682	0.620	0.611
1,500	0.704	0.682	0.616	0.607

training partitions of individual folds, we applied Pearson's correlation (PeaCorr) tests to reduce the feature dimension from 9910 to $\{250, 500, 750, 1000, 1500\}$, as shown in Table 2. By reducing the feature dimension to 500, we obtained the best CV performance on the training data. Therefore, on the training partitions of individual folds, subsequent experiments utilized PeaCorr to reduce the feature dimension to 500. On the pre-screened features, we further applied DDR to select discriminative features. We followed [15] and obtained the optimal feature subsets by varying the number of selected features through CV. The optimal feature subset was obtained when the highest F_1 score was achieved in the CV, which is 18, as shown in Table 1. DDR significantly reduces the feature dimensions and achieves the best recognition performance on the training data.

4.2. Evaluation on Text Embeddings

When fine-tuning the ELECTRA model, we determined the best possible hyper-parameter settings using grid-search and CV. The evaluation results are shown in Table 3, which shows that concatenating the text embeddings with confidence scores can substantially improve performance. When encoded with DP, recognition performance is further improved. We depict the distributions of β/α (when $\alpha > 0.25$ sec) in the dementia patients and HCs in Figure 3. Figure 3 clearly shows that the dementia patients have more β/α in the interval $[0, 3)$ than the HCs.

4.3. Fusing Two Branches

The selected features from Branch 1 were used to obtain recognition results on test data. At the same time, we selected the best model from Branch 2 to obtain recognition results on test data. Finally, we fused the two recognition results by averaging the scores from the two branches, as shown in Table 4. It shows that on the test data, the recognition performance of Branch 2 is significantly better than Branch 1. When fusing results from the two branches, even better performance on the test data has been achieved.

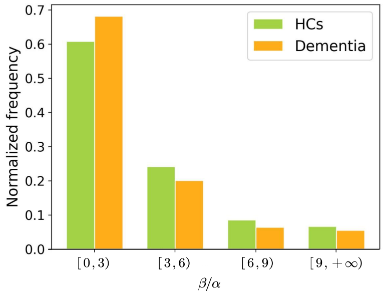

Fig. 3. The distributions of β/α (when $\alpha > 0.25$ sec) in HCs and dementia patients.

Table 3. Classification performance on text embeddings. CS: confidence score; DP: disfluency pattern; *e*=epochs; *mwl*=max word length; *bs*=batch size.

Methods	**10-fold CV**		**Parameters**
	ACC	F_1	
Multiple hypotheses	0.651	0.612	*e*=4, *mwl*=128, *bs*=32
Multiple hypotheses + CS	0.701	0.645	*e*=4, *mwl*=128, *bs*=32
Multiple hypotheses + DP	0.665	0.627	*e*=4, *mwl*=128, *bs*=32
Multiple hypotheses + CS + DP	***0.707***	***0.646***	*e*=4, *mwl*=128, *bs*=32, *a*=0.25 sec, *b*=3.0

Table 4. Final classification results on the test data.

Method	**Performance on test data**			
	ACC	PRE	REC	F_1
PeaCorr + DDR (Branch 1)	0.708	0.737	0.708	0.699
Multiple hypotheses + CS + DP (Branch 2)	0.742	0.754	0.742	0.739
Branch 1 + Branch 2	***0.750***	***0.764***	***0.750***	***0.747***

5. CONCLUSIONS

The Hubert model was used to transcribe the Marvel Cantonese corpus. We presented an end-to-end system containing two branches and evaluated the system on the Marvel dataset for dementia detection. We analyzed and utilized disfluency patterns to improve detection performance. The combination of the two branches improves the performance on the Marvel test data.

6. ACKNOWLEDGMENTS

This work was in part supported by Research Grants Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N).

7. REFERENCES

[1] J. L. Cummings, R. Doody, and C. Clark, "Disease-modifying therapies for Alzheimer disease: Challenges to early intervention," *Neurology*, vol. 69, no. 16, pp. 1622–1634, 2007.

[2] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dement. Geriatr. Cogn. Disord.*, vol. 37, no. 5-6, pp. 327–334, 2014.

[3] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 272–281, 2020.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[5] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia International Conference*, 2010, pp. 1459–1462.

[6] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagramcochleagram features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 3737–3741.

[7] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, "Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models," in *Proc. Interspeech*, 2021, pp. 3805–3809.

[8] M. Ströbel, E. Kerz, and D. Wiechmann, "The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning," *Lang. Learn.*, vol. 70, no. 3, pp. 732–767, 2020.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Porc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[10] X. Lu and J. Wu, "Noun-phrase complexity measures in Chinese and their relationship to l2 Chinese writing quality: A comparison with topic–comment-unit-based measures," *The Modern Language Journal*, vol. 106, no. 1, pp. 267–283, 2022.

[11] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint arXiv:cs/0205028v1*, 2002.

[12] M. Komeili, C. Pou-Prom, D. Liaqat, K. C. Fraser, M. Yancheva, and F. Rudzicz, "Talk2me: Automated linguistic data collection for personal assessment," *PLoS One*, vol. 14, no. 3, p. e0212342, 2019.

[13] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — a collaborative voice analysis repository for speech technologies," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 960–964.

[14] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, 2010, pp. 2794–2797.

[15] X. Ke, M. W. Mak, J. Li, and H. M. Meng, "Dual dropout ranking of linguistic features for Alzheimer's disease recognition," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021, pp. 743–749.

[16] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the outputs of different automatic speech recognition paradigms for acoustic- and bert-based Alzheimer's dementia detection through spontaneous speech." in *Proc. Interspeech*, 2021, pp. 3810–3814.

[17] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease." in *Proc. Interspeech*, 2020, pp. 2162–2166.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Neural Information Processing Systems*, 2020, pp. 12 449–12 460.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *arXiv preprint arXiv:2106.07447*, 2021.

[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. Conference on Language Resources and Evaluation*, 2020, pp. 4211–4215.