Jointly Modelling Transcriptions and Phonemes with Optimal Features to Detect Dementia from Spontaneous Cantonese

Xiaoquan Ke*, Man-Wai Mak*, and Helen M. Meng†
* The Hong Kong Polytechnic University, Hong Kong SAR
† The Chinese University of Hong Kong, Hong Kong SAR
xiaoquan.ke@connect.polyu.hk, enmwmak@polyu.edu.hk, hmmeng@se.cuhk.edu.hk

***Abstract*—Dementia is a severe cognitive impairment that affects the health of older adults. This paper analyzes diverse speech-based features extracted from spoken languages and selects the most discriminative ones for dementia detection. We propose a two-step feature selection method to handle the circumstance where the feature dimension is far larger than the number of training samples. Recently, the performance of dementia detection has been significantly improved by utilizing Transformer-based models that automatically capture the linguistic properties of spoken languages. We combine features extracted from BERT with selected speech-based features to enhance dementia detection performance. We propose a novel strategy to model the transcriptions and their phonemes using BERT and phoneme-BERT. The proposed method is evaluated on a Cantonese dataset called CU-Marvel, which contains 185 healthy older adults, 98 older adults having minor neurocognitive disorders (minor NCD), and 26 older adults suffering from major NCD. Experimental results show that simultaneously fine-tuning the BERT and phoneme-BERT can leverage information from the recognized phonemes and make the detection performance robust to automatic speech recognition errors. Simultaneous fine-tuning of the BERT and phoneme-BERT models results in a 6% improvement in F_1 scores, compared to fine-tuning the BERT model alone.**

Index Terms: Dementia detection, feature selection, phoneme-BERT

I. Introduction

Dementia is a severe cognitive impairment that may seriously affect the health and daily lives of the afflicted individuals. The most common form of dementia is Alzheimer's Disease (AD). According to a report from the World Health Organization,[1] more than 55 million people live with dementia worldwide, and there are nearly 10 million new cases every year. 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].

A. Related Works

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, Haider *et al.* [2] compared different types of paralinguistic features – including the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [3], ComParE 2013 [4], Emobase [4], and MRCG [5] – for dementia detection. Because 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 [6–8]. For example, Qiao *et al.* [8] combined disfluency and linguistic complexity features with features extracted from Transformer-based models for AD detection. In [8], BERT [9] and ERNIE [10] models were fine-tuned to capture the language characteristics of the AD patients. The Transformer-based models were also extensively investigated by Syed *et al.* [6].

More recently, the fully-automatic assessment of dementia from spontaneous speech has gathered more attention as it does not require labor-intensive annotations or manual transcriptions. In such a case, an automatic speech recognition (ASR) system is utilized to transcribe the patients' speech. However, the erroneous transcriptions produced by the ASR system could lead to high detection errors. To mitigate the errors in ASR systems, the research community has developed three strategies:

- *Adapting ASR systems.* In the Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge [11], the conventional Kaldi-based ASR system was adapted using multiple datasets containing spontaneous speech [12]. Pappagari *et al.* [13] first used a pre-trained ASpIRE model from Kaldi to transcribe target domain data. To improve transcription quality, they interpolated the language model (LM) of ASpIRE with an LM trained on automatic transcriptions of the target domain.
- *Utilizing ASR lattices.* An ASR lattice can provide time alignments, recognized words, and confidence scores for different hypotheses. Usually, we only consider the best hypothesis that has the highest confidence score.

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

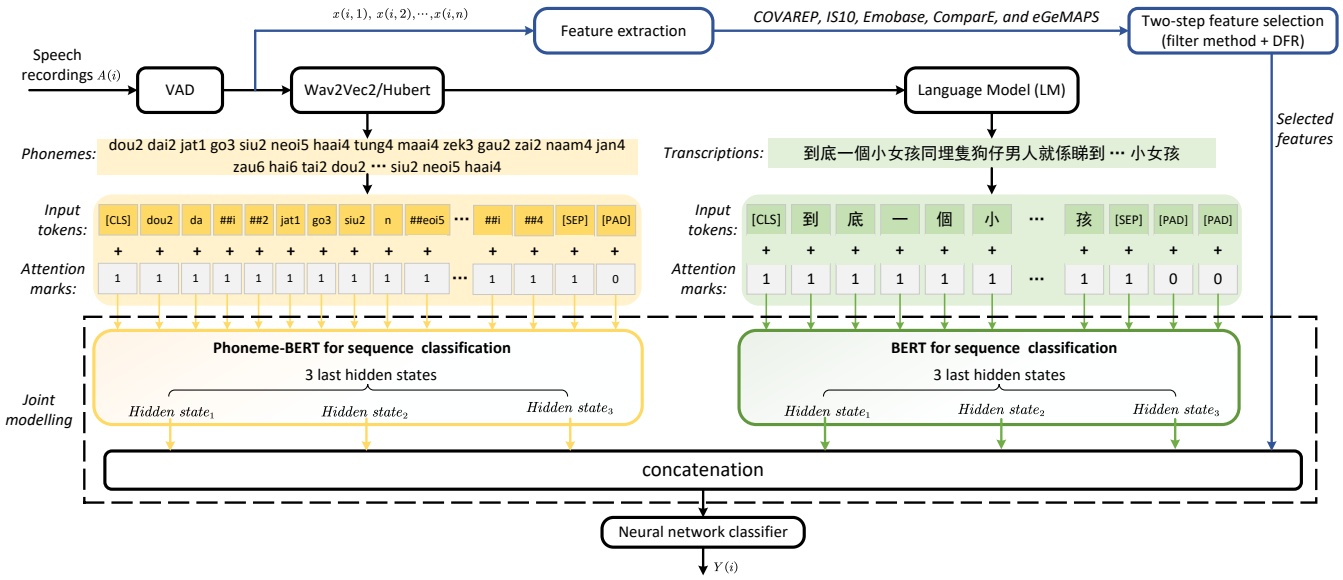

Fig. 1. The architecture of our system for detecting dementia of Cantonese speakers. The recording $A(i)$ is segmented using voice activity detection into n segments. A two-step FS method identifies the most discriminative speech-based features. BERT and phoneme-BERT are jointly fine-tuned on ASR transcriptions and their corresponding phonemes and are combined with the speech-based features to screen dementia patients. We concatenated the last three hidden states of the BERT model, the last three hidden states of the phoneme-BERT model, and the selected features. The neural network classifier was adopted for classification.

However, in addition to the best hypothesis, Pan *et al.* [12] used multiple hypotheses to augment the input of BERT to recognize AD patients. More importantly, they concatenated the confidence scores and hidden states of the BERT model, where the confidence scores were used as a proxy measure of accuracy to inform the classifier about the transcription quality.

- *ASR correction.* To extract more robust linguistic features to distinguish AD patients, the transcribed words with low confidence scores can be removed from the automatic transcriptions [14]. In [12], the automatically transcribed words with the confidence scores lower than 0.87 were replaced by the unknown token <unk>.

B. Modeling Approach

The modeling approach presented in this paper is based on the key insights from the above studies. Importantly, it combines the most discriminative speech-based features and the features extracted from Transformer-based models for fully-automatic dementia screening. To overcome the data sparsity problem, we propose a two-step feature selection (FS) method to identify the most discriminative speech-based features. Additionally, to mitigate the errors in ASR systems, we propose a new strategy to jointly model the ASR transcripts and their corresponding phonemes using the Transformer-based models. The whole architecture is shown in Fig. 1. The proposed system is evaluated on a Cantonese corpus called CU-Marvel. The main contributions of this work are summarized as follows:

1) We propose a two-step FS method to identify the most discriminative speech-based features to screen dementia patients.
2) Extensive experiments are conducted to explore the possibilities of modelling phonemes using phoneme-BERT to detect dementia. Results show that information underlying the ASR phonemes is also indicative of dementia.
3) We jointly model transcriptions and phonemes using BERT and phoneme-BERT. Experimental results show that simultaneously fine-tuning the BERT and phoneme-BERT models can leverage information from the recognized phonemes, making the dementia classifier more robust to ASR errors.

II. Methods

A. Selection of Speech-based Features

While various types of 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 FS methods to find the most effective speech-based features for dementia detection. The extracted speech-based features are as follows.

- *INTERSPEECH 2010 Paralinguistic Challenge Features (IS10).* IS10 [15] is a feature set for emotion recognition and bipolar disorder recognition. In addition to the 32 low-level descriptors (LLDs) in INTERSPEECH 2009 Emotion Challenge, 44 LLDs were added to IS10, including PCM loudness, eight log Mel-frequency bands, eight line-spectral frequency pairs, fundamental frequency (F0) envelope, voicing probability, jitter, and shimmer. Twelve statistics (minimum, maximum, mean, range, etc.) of the LLDs were computed, leading to a 1582-dimensional feature vector per recording. IS10 was adopted as the

baseline feature set for the AD2021 Alzheimer's Dementia Recognition Challenge.[2]

- *COVAREP.* COVAREP [16] provides comprehensive acoustic features, which include prosodic features (F0 and voicing), voice quality features, and spectral features. We extracted COVAREP features at 100Hz; for each recording, the mean, maximum, minimum, median, standard deviation, skew, and kurtosis of the features were computed, leading to a 518-dimensional feature vector per recording. Rohanian *et al.* [17] used the COVAREP features for cognitive impairment detection.
- *eGeMAPS.* The eGeMAPS [3] contains 88 features that are selected based on their potential for characterizing physiological changes in voice production.
- *ComParE 2013.* The ComParE 2013 [4] feature set was adopted as a baseline feature set for AD detection in the ADReSS [18] and ADReSSo [11] challenges, which contains energy, spectral, mel-frequency cepstral coefficients (MFCC), and voicing related LLDs.
- *Emobase.* The Emobase feature set [4] comprises MFCC, F0, F0 envelope, line spectral pairs, etc. Wang *et al.* [19] used the Emobase feature set in multi-modal attention network for AD detection.
- *MRCG.* The MRCG features are multi-resolution cochleagram features, which were used by Haider *et al.* [2] to identify AD patients.

We combined all the feature listed above and adopted a two-step FS approach to selecting dementia features. When the feature dimension is very high, filter methods are indispensable for obtaining a reduced feature set for the expensive FS methods. Therefore, in Step 1, filter methods are utilized to pre-screen the original features. Three filter methods were evaluated in the experiments: Fisher's discriminant ratio (FDR) [20], PeaCorr tests, and mutual information (MutInfo). In Step 2, a deep-learning-based feature ranking method called dual-net feature ranking (DFR) [21] is applied to rank the remaining features. Features with high feature importance were then selected. DFR utilizes a dual-net architecture, where two networks (called operator and selector) are trained to simultaneously perform FS and dementia detection. Specifically, the selector is trained to find multiple subsets of features with minimal cardinality to predict the operator's performance, and the operator uses these feature subsets to minimize classification errors. DFR uses all of the selector's parameters to determine the contribution of individual features to the selector's predictions. DFR shows good performance on a dementia-related Cantonese dataset called JCCOCC-MoCA [22].

B. Automatic Speech Recognition

In order to build a fully-automatic dementia screening system, ASR was used for transcribing speech. Wav2vec 2.0 [23] (denoted as Wav2vec2 from now on) and HuBERT [24] are self-supervised pre-trained models that can be utilized for end-to-end ASR. They can also learn powerful representations from a large amount of unlabeled speech data. By fine-tuning the models on a small amount of transcribed speech, they can achieve similar performance as the traditional fully-supervised ASR systems [23]. 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*,[3] *Wav2vec2-large-Chinese*,[4] and *Hubert-large-Chinese*.[5]

The Cantonese version of Common Voice Speech dataset [25] (common-voice-zh-HK) was used for fine-tuning. The PyCantonese library was utilized to convert the transcriptions to corresponding phonemes.[6] The acoustic models were end-to-end fine-tuned on phone-level using connectionist temporal classification loss. The fine-tuned acoustic models were tested on common-voice-zh-HK test data. The phone error rate for *Wav2vec2-large-xlsr*, *Wav2vec2-large-Chinese*, and *Hubert-large-Chinese* were 0.087, 0.145, and 0.148, respectively. Therefore, *Wav2vec2-large-xlsr* was selected for producing phonemes using the CU-Marvel corpus as inputs. To generate the ASR transcriptions, the phonemes were decoded using a beam search decoder with a 4-gram KenLM trained on common-voice-zh-HK.[7]

C. Modelling with Optimal Features

The BERT models were pre-trained on a large amount of transcriptions using masked language modeling (MLM) as the pre-training objective. As there is no pre-trained version of BERT on Cantonese phonemes, we pre-trained our phoneme-BERT using the common-voice-zh-HK Cantonese phonemes. To form a better representation of the phonemes, we first trained a Word Piece Encoder (WPE) with a vocabulary that consists of 600 sub-word units[8] using the common-voice-zh-HK phonemes. The WPE was shown to work well on the Cantonese phonemes. The trained WPE was then utilized to encode the phonemes into multiple WPE-tokens. In our pre-training settings, the token embeddings of WPE-tokens were randomly initialized and the default configuration of the original BERT was adopted. Following the pre-training regime of BERT, we randomly masked 20% of WPE-tokens using `[MASK]`. Finally, we pre-trained our phoneme-BERT using the masked WPE-tokens as inputs and MLM as the learning objective.

After pre-training our phoneme-BERT, we combined it with a well pre-trained BERT model[9] from the Transformer library. We followed the strategy in [12] and concatenated the last three hidden states of the BERT model with the last three hidden states of the phoneme-BERT model, see

[2]https://github.com/THUsatlab/AD2021

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

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

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

[6]https://pycantonese.org/

[7]Decoding parameters: language model's weight $= 3.0$, word score $= 0.0$, number of best hypotheses $= 1$, and beam size $= 500$.

[8]We set the size of the vocabulary to 600 according to training procedure of a byte pair encoder in [26].

[9]https://huggingface.co/bert-base-chinese

TABLE I
CHARACTERISTICS OF THE CU-MARVEL DATASET.

Variable	HCs ($n = 185$)	Dementia ($n = 124$)
No. of Female	122	70
No. of Male	63	54
Age (years)	70.0 (66.0, 75.0)	78.0 (68.8, 83.0)
Education (years)	7.0 (6.0, 11.0)	6.0 (2.0, 9.25)
MoCA scores	23.0 (21.0, 26.0)	19.0 (15.0, 21.0)
Language	Cantonese	
Task	Rabbit-story picture description	
Manual transcriptions	No	

HCs: healthy older adults; MoCA: Montreal Cognitive Assessment.
The values are presented as *median (interquartile range)*.

Fig. 1. To utilize information from the speech, we combined the concatenated hidden states with the selected speech-based features. Note that the selected speech-based features were normalized to make them compatible with the characteristics of the hidden states. A fully-connected neural network (FCNN) classifier was built on top of the concatenated hidden states and the selected features. Finally, the phoneme-BERT and BERT were end-to-end fine-tuned on the ASR transcriptions and their corresponding phonemes.

III. DATASETS AND EXPERIMENTAL SETTINGS

A. The CU-Marvel Cantonese Corpus

Cantonese is one of the major Chinese dialects that has over 80 million native speakers in Southern China. The CU-Marvel corpus was collected for the research on the screening and monitoring of neurocognitive disorders based on spoken language technologies. 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 and dementia in older adults. According to the assessment results, 309 participants were divided into three groups: 185 healthy older adults (HCs), 98 older adults having minor neurocognitive disorders (minor NCD), and 26 older adults suffering from major NCD. For detecting dementia, we combined minor NCD and major NCD into one category called "possible dementia". A rabbit story picture description task was selected for the experiments. Table I shows the corpus's characteristics.

B. Implementation Details

The phoneme-BERT was pre-trained on the common-voice-zh-HK phonemes with a batch size of 512 for 3,000 steps. We observed an increase in the validation loss when the number of steps exceeded 3,000. 10-fold cross-validation (CV) was applied to the CU-Marvel dataset to determine the performance of dementia detection. For each fold, we fine-tuned the models with a batch size of 4 for 10 epochs. The maximum length for word tokens and phoneme tokens was set to 512. This setting aims to cover as many words and phonemes as possible because most participants spoke a lot.

TABLE II
DETECTION PERFORMANCE (F_1 SCORES) ON THE CU-MARVEL. THE NUMBERS IN THE BRACKETS ARE FEATURE DIMENSIONS.

Feature set	SVM	DT	KNN	LDA	FCNN
IS10 (1582)	0.521	0.559	0.528	0.517	0.590
COVAREP (518)	0.567	0.552	0.564	0.562	0.574
eGeMAPS (88)	0.571	0.582	0.525	0.539	0.614
ComparE (6373)	0.556	0.526	0.555	0.569	0.603
Emobase (988)	0.497	0.540	0.521	0.503	0.630
MRCG (768)	0.529	0.533	0.543	0.495	0.605
Mean	0.540	0.549	0.539	0.531	***0.603***

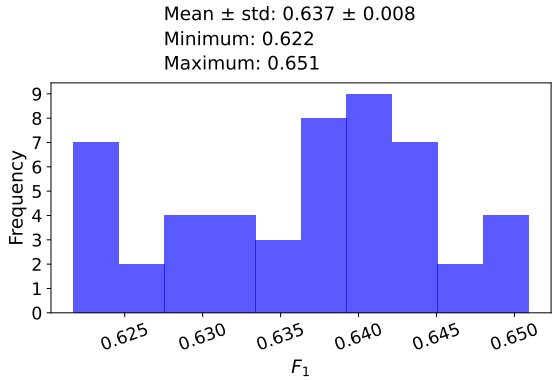

Fig. 2. Variation in detection performance across the repeated CVs.

IV. RESULTS AND DISCUSSIONS

A. Performance of Different Feature Types

We first evaluated the recognition performance of all the feature sets *before* FS. Five classifiers were adopted, including k-nearest neighbor (k-NN) classifier ($k = 1$), linear support vector machines (SVM, box constraint $= 0.1$), decision trees (DT, leaf size $= 20$), linear discriminant analysis (LDA), and FCNN classifier.[10] We ran 50 repetitions of the 10-fold CV based on different data splittings and averaged the performance (F_1 scores). The corresponding results are reported in Table II. Table II shows that the FCNN classifier achieves the best classification performance among all classifiers, that is, it achieves the highest averaged F_1 scores. Therefore, subsequent experiments adopted the FCNN classifier.

We then combined all of the features to form the *combined features*, which are 10317-dimensional vectors. When conducting 10-fold CV on the combined features, large variation in detection performance across the repeated CVs was observed, as illustrated in Fig. 2. The variation is caused by applying random splitting on a small dataset for each CV, inducing significantly different training partitions (TR) for each run. To reduce the variation, we propose using an ensemble procedure to stabilize the detection performance. Specifically, we ran 10 repetitions of CV based on different data splittings.

[10]Except for the FCNN classifier, all other classifiers were adopted from [2]. FCNN classifier settings: network architecture 'feature_dimension–128–32–2", batch size $= 4$, and epochs $= 10$.

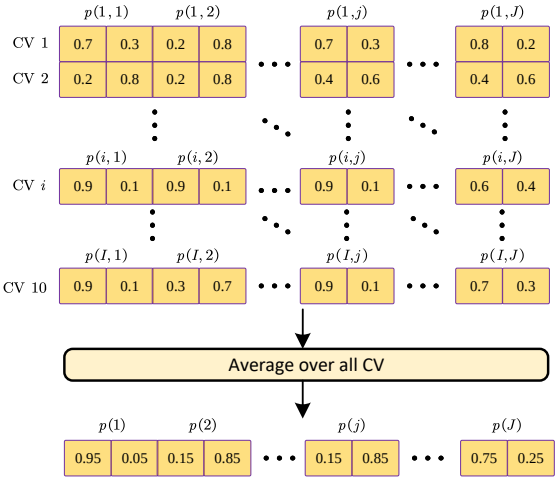

Fig. 3. The ensemble procedure to stabilize the classification performance estimated by CV. We ran I repetitions of CV based on different data splittings and averaged the predicted scores $p(i,j)$ over all the CV for each of the J subjects.

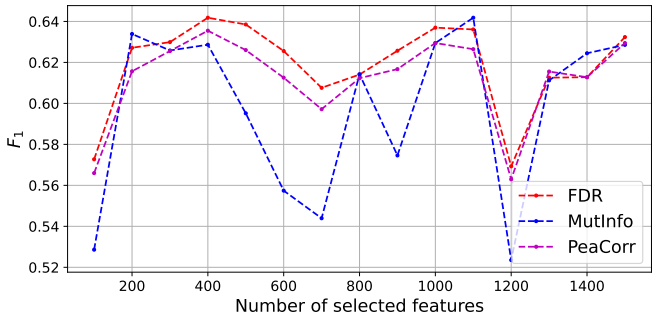

Fig. 4. Classification performance of the filter methods.

We then produced the predicted scores $p(i,j)$ for subject j in CV i, as shown in Fig. 3. Finally, we averaged the predicted scores $p(j) = (1/10)\sum_{i=1}^{10} p\,(i,j)$ over all the 10 CV for each of the J subjects. After utilizing the ensemble procedure, we obtained 0.670 accuracy (ACC) and 0.645 F_1 scores, which boosts the minimum F_1 scores from 0.622 (Fig. 2) to 0.645. Therefore, the subsequent experiments adopted the ensemble procedure to stabilize classification performance.

B. Performance of Two-step FS

We followed the procedure described in Section II-A to evaluate the performance of the filter methods (FDR, PeaCorr, and MutInfo) on the combined features (10317-dimensional vectors). On the TR of individual folds, we applied the filter methods to reduce the feature dimension to $n = \{100, 200, 300, 400, \ldots, 1500\}$, as shown in Fig. 4. It shows that FDR achieves the highest F_1 scores (0.645) when the feature dimension was reduced to 400. We then applied DFR on the remaining 400 features to further select more discriminative features, as shown in Table III. It shows that when we reduced the feature dimension from 400 to 300, we further improved the ACC. This two-step procedure

TABLE III
CLASSIFICATION PERFORMANCE OF TWO-STEP FS. *ACC*: ACCURACY; *PRE*: PRECISION; *REC*: RECALL.

	Dimension	10-fold CV ACC	PRE	REC	F_1
FDR + DFR (Step 1 + Step 2)	150	0.650	0.634	0.631	0.632
	200	0.663	0.648	0.626	0.627
	250	0.628	0.623	0.628	0.622
	300	***0.680***	***0.668***	***0.643***	***0.645***
	350	0.634	0.622	0.624	0.623
	400	0.663	0.648	***0.643***	***0.645***

significantly reduces the feature dimension and outperforms the combined features.

C. Performance of Fine-tuning

We evaluated the performance on fine-tuning BERT, phoneme-BERT, and the combination of BERT and phoneme-BEET, as shown in Table IV (Row 2 to Row 4). Row 3 shows that fine-tuning phoneme-BERT on ASR phonemes is beneficial to dementia detection. This indicates that some abnormal information (e.g., phoneme disorder, phoneme repetition, phoneme scarcity, and phoneme deficiency) in the patients' ASR phonemes is also indicative of dementia. Row 2 and Row 3 show that fine-tuning phoneme-BERT is better than fine-tuning BERT. This maybe due to the insufficient training of the LM. If we adapt the LM with an elderly corpus containing similar picture description tasks, we may reduce the ASR errors and improve detection performance.

Row 4 shows that it is better to fine-tune the phoneme-BERT and BERT models jointly instead of fine-tuning the BERT model only, indicating that the recognized phonemes contain useful information for dementia detection. Our ASR system produces phonemes, which are subsequently decoded into words using a 4-gram KenLM. In Cantonese, a single set of phonemes may correspond to multiple words, i.e., homophones. Despite the ASR system producing the correct phonemes, in some instances, these phonemes are misinterpreted, resulting in erroneous words. Though the ASR system decodes some erroneous words, simultaneous fine-tuning of BERT and phoneme-BERT enables the system to leverage the corresponding correct phonemes, enhancing its overall robustness to ASR errors.

D. Fine-tuning with Optimal Features

Finally, we fine-tuned BERT, phoneme-BERT and the combination of phoneme-BERT and BERT with the optimal features selected in Section IV-B. Table IV (Row 5 to Row 7) shows that fine-tuning with the optimal features can improve performance, no matter on which model. This strategy leverages the optimal features to enhance the inputs to the classifier. Row 7 shows that with the optimal feature set, simultaneously fine-tuning the phoneme-BERT and BERT models achieves the best detection performance.

TABLE IV
CLASSIFICATION PERFORMANCE ACHIEVED BY FINE-TUNING DIFFERENT MODELS. *ACC*: ACCURACY; *PRE*: PRECISION; *REC*: RECALL.

Row		**Fine-tuning**		**10-fold CV**			
	FDR + DFR	BERT	Phoneme-BERT	ACC	PRE	REC	F_1
1	✔	✘	✘	0.680	0.668	0.643	0.645
2	✘	✔	✘	0.608	0.591	0.591	0.591
3	✘	✘	✔	0.650	0.635	0.634	0.634
4	✘	✔	✔	0.667	0.654	0.655	0.654
5	✔	✔	✘	0.693	0.680	0.666	0.669
6	✔	✘	✔	0.683	0.670	0.671	0.671
7	✔	✔	✔	***0.696***	***0.685***	***0.687***	***0.686***

V. CONCLUSIONS

We proposed a two-step feature selection method to identify the most effective speech-based features to screen Cantonese-speaking dementia patients. Fine-tuning Transformer-based models with the selected speech-based features improved performance further. We also fine-tuned BERT and phoneme-BERT on transcriptions and their corresponding phonemes, which made the dementia classifier more robust to ASR errors. Fine-tuning both the BERT and phoneme-BERT models together resulted in a 6% improvement in F_1 scores, compared to fine-tuning the BERT model alone.

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, Oct. 2007.

[2] 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, Feb. 2020.

[3] 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, Apr. 2016.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia Int. Conf.*, Oct. 2010, pp. 1459–1462.

[5] F. Haider and S. Luz, "Attitude recognition using multi-resolution Cochleagram features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2019, pp. 3737–3741.

[6] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Tackling the ADRESSO challenge 2021: The MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech*, Aug. 2021, pp. 3815–3819.

[7] J. Li, J. Yu, Z. Ye, S. Wong, M. W. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6423–6427.

[8] 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*, Aug. 2021, pp. 3805–3809.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805

[10] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 8968–8975.

[11] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," in *Proc. Interspeech*, Aug. 2021, pp. 4211–4215.

[12] 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*, Aug. 2021, pp. 3810–3814.

[13] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios," in *Proc. Interspeech*, Aug. 2021, pp. 3825–3829.

[14] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *Proc. Interspeech*, Oct. 2020, pp. 159–164.

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

[16] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.

[17] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," 2021, *arXiv:2106.09668*. [Online]. Available: https://doi.org/10.48550/arXiv.2106.09668

[18] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," 2020, *arXiv:2004.06833*. [Online]. Available: https://arxiv.org/abs/2004.06833

[19] N. Wang, Y. Cao, S. Hao, Z. Shao, and K. Subbalakshmi, "Modular multi-modal attention network for Alzheimer' disease detection using patient audio and language data," in *Proc. Interspeech*, Aug. 2021, pp. 3835–3839.

[20] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8696–8702, Jul. 2011.

[21] X. KE, M.-W. Mak, and H. M. Meng, "Automatic selection of discriminative features for dementia detection in Cantonese-speaking people," in *Interspeech 2022*, Sep. 2022, pp. 2153–2157.

[22] S. S. Xu, M. W. Mak, K. H. Wong, H. Meng, and C. Y. Kwok, "Speaker turn aware similarity scoring for diarization of speech-based cognitive assessments," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC*. IEEE, 2021, pp. 1299–1304.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. neural inf. proces. syst. (NIPS)*, Dec. 2020, pp. 12 449–12 460.

[24] 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," 2021, *arXiv:2106.07447*. [Online]. Available: https://doi.org/10.48550/arXiv.2106.07447

[25] 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," 2020, *arXiv:1912.06670*. [Online]. Available: https://doi.org/10.48550/arXiv.1912.06670

[26] M. N. Sundararaman, A. Kumar, and J. Vepa, "PhonemeBERT: Joint language modelling of phoneme sequence and ASR transcript," in *Proc. Interspeech*, Aug. 2021, pp. 3236–3240.