Dementia Detection by Analyzing Spontaneous Mandarin Speech

Zhaoci Liu*, Zhiqiang Guo*, Zhenhua Ling*, Shijin Wang†, Lingjing Jin‡ and Yunxia Li‡
* University of Science and Technology of China, Hefei, China
E-mail: zcliu8@mail.ustc.edu.cn, gzq@mail.ustc.edu.cn, zhling@ustc.edu.cn
† iFLYTEK Research, iFLYTEK Co., Ltd., Hefei, China
State Key Laboratory of Cognitive Intelligence, Hefei, China
E-mail: sjwang3@iflytek.com
‡ Shanghai Tongji Hospital, Tongji University School of Medicine, Shanghai, China
E-mail: lingjingjin@163.com, doctorliyunxia@163.com

***Abstract*—The Chinese population has been aging rapidly resulting in the largest population of people with dementia. Unfortunately, current screening and diagnosis of dementia rely on the evidences from cognitive tests, which are usually expensive and time consuming. Therefore, this paper studies the methods of detecting dementia by analyzing the spontaneous speech produced by Mandarin speakers in a picture description task. First, a Mandarin speech dataset contains speech from both healthy controls and patients with mild cognitive impairment (MCI) or dementia is built. Then, three categories of features, including duration features, acoustic features and linguistic features, are extracted from speech recordings and are compared by building logistic regression classifiers for dementia detection. The best performance of identifying dementia from healthy controls is obtained by fusing all features and the accuracy is 81.9% in a 10-fold cross-validation. The importance of different features is further analyzed by experiments, which indicate that the difference of perplexities derived from language models is the most effective one.**

***Index Terms*—Alzheimer's disease, dementia detection, speech analysis, logistic regression, language model.**

I. Introduction

Dementia is a neurodegeneration disorder that develops for years , with Alzheimer's disease (AD) being the most common underlying pathology [1], [2]. The course of dementia can be divided into four stages, i.e., mild cognitive impairment (MCI) stage, early stages, middle stage and late stage, according to the progressive degree of cognitive and functional impairment. During this process, patients suffer from short memory loss at the beginning and completely depend upon caregivers at last. The Chinese population has been aging rapidly resulting in the largest population of people with dementia. Statistics show that the prevalence of dementia among individuals aged 65 years and older were 5.14% in China [3] . The estimated total annual costs of dementia in China were US$47.2 billion in 2010 and were predicted to reach US$69.0 billion in 2020 and US$114.2 billion in 2030 [4].

Current diagnosis of dementia relies on the evidences from cognitive tests, biochemical markers, medical imaging, etc., which are usually expensive and time consuming. Furthermore, there is no cure for dementia so far. Thus, it is valuable if some low-cost and convenient detection methods can be developed to find the dementia patients at their early-stage for proper prevention and intervention therapies.

Language impairment is one of the main symptoms of dementia, which generally appears at the early stages of the disease [5], [6]. Some investigations [7]–[10] found that AD patients suffered from word finding and word retrieval difficulties. Their performances on some cognitive tasks, such as picture description and sentence repetition, were distinct from healthy people. Thus, the effective detection of dementia can be achieved by extracting proper features from speech recordings and building classifiers in a data-driven way.

Several databases for studying the speech and language impairment of dementia patients have already been established. DementiaBank Pitt corpus [11] contained the recordings from 312 English speakers taking a picture description task. Most of previous studies using this dataset [12]–[15] aimed to make a binary classification between AD group and control group, which had about 250 and 240 samples in the dataset. Fraser et al. [12] extracted total 370 features considering part-of-speech (POS), syntax, acoustics and other aspects of linguistics, and obtained an accuracy of 81% in binary classification. Warnita et al. [13] used a gated convolutional neural network (GCNN) which utilized only the audio data and achieved an accuracy of 73.6%. Wankerl et al. [14] and Fritsch et al. [15] calculated the difference of perplexities from the language models of two groups. Using this single feature, their methods achieved an accuracy of 77.1% and 85.6% at equal-error-rate respectively.

A Germany database named *the interdisciplinary longitudinal study on adult development and aging* (ILSE) was created by Weiner et al. [16]. In their latest work, they utilized 98 recordings from 74 recruited people. The participants were divided into three group, including healthy controls, the ones with aging-associated cognitive decline (AACD) and the ones with Alzheimers disease (AD). The participant's speech was recorded in biographic interviews. In the latest experiments, they extracted features from the conversational speech through voice activity detection (VAD) and speaker diarization. Their three-way classifier achieved the average recall (UAR) of 0.645 [17].

The Hungarian MCI-mAD database was built by Hoffmann

et al. [18]. As introduced in their recent work [19], there were 75 speakers and 225 recordings captured form three tasks, *immediate recall*, *previous day* and *delayed recall*. They obtained an accuracy of 80% when identifying MCI and mild AD using linguistic features.

Satt et al. [20] recruited 15 health controls and 26 patients suffer from dementia in France and collected their recordings of 4 cognitive tests. They obtained an equal error rate (EER) of 87% for binary classification using carefully designed acoustic features for different tasks. However, this result lacked large-scale verification.

To the best of our knowledge, there are no existing large-scale Mandarin speech dataset for developing and verifying dementia detection models so far. Therefore, we first constructs a dataset containing spontaneous speech produced by Mandarin speakers in a picture description task. The speakers include healthy controls and patients with mild cognitive impairment (MCI) or dementia. Then, this paper focuses on the task of identifying dementia from healthy controls. Three categories of features, including duration features, acoustic features and linguistic features, are extracted and logistic regression classifiers are built using these features. After fusing the features of all categories, the dementia detection accuracy of 81.9% is finally obtained. We also analyze the importance of different feature categories by examining their weights in logistic regression and their area under the curve (AUC) values. The results show that linguistic features play the most important role in our model.

II. Dataset

A. Subjects

Subjects were recruited from the Department of Neurology and the Department of Memory Clinic of Shanghai Tongji Hospital. All participants were with the complaint of memory impairment and underwent a comprehensive neuropsychological battery that included the Mini-Mental State Examination (MMSE) [21], the Chinese version of the Montreal Cognitive Assessment Basic (MoCA-BC) [22], the Clinical Dementia Rating (CDR) [23], the Instrumental Activities of Daily Living(IADL) [24], the Hopkins Verbal Learning Test-Revised (HVLT-R) [25], the Shape Trail Test-A and B (STT-A, STT-B) [26], the Boston Naming Test (BNT; the 30-item version) [27], the Rey-Osterrieth Complex Figure Test (CFT) [28], the Hamilton Depression Rating Scale [29] and the picture description task. Speech recordings were collected at the same time. All participants underwent cranial CT or MRI scan and laboratory screening on folic acid, vitamin B12, thyroid function (free triiodothyronine(FT3), free tetraiodothyronine(FT4), thyroid stimulating hormone(TSH)), treponema pallidum and HIV antibodies. Their demographic and clinical information was recorded at the same time. Exclusion criteria: 1) age below 40 years; 2) less than 5 years of education; 3) definite history of stroke; 4) definite history of other diseases of the central nervous system such as infection, demyelinating diseases, and Parkinson's disease; 5) definite history of mental illness such as schizophrenia, major depressive disorder; 6) serious physical disease; 7) alcohol or drug addiction; 8) with clinically significant abnormalities in folic acid, vitamin B12, thyroid function, or syphilis antibody positive, HIV antibody; 9) unable to cooperate with neuropsychological tests. Written informed consents were obtained from all participants. Finally, the participants (also referred as *subjects*) were categorized in three groups.

- **Dementia** – Dementia diagnosis was made according to the core clinical criteria to dementia of NIA-AA established in 2011 [30]. And three categories of dementia patients were included: dementia due to Alzheimer's disease, dementia due to cerebral small vessel disorder, Alzheimer's disease mixed with cerebral small vessel disorder.
- **MCI** – the participants who were diagnosed as MCI. MCI diagnosis was made according to the guidelines of NIA-AA established in 2011 [30].
- **CTRL** – the participants who joined the cognitive test but were diagnosed as cognitively healthy.

TABLE I
THE STATISTICS OF THE SUBJECTS IN OUR DATASET.

Group (number)	Gender (Male/ Female)	Age mean (std)	Education mean (std)	MoCA-BC mean (std)	MMSE mean (std)
CTRL (138)	59/79	66.6 (9.3)	11.8 (3.0)	23.3 (3.0)	27.8 (1.6)
MCI (179)	71/108	66.0 (9.6)	10.5 (2.9)	18.0 (4.4)	24.9 (3.4)
Dementia (84)	39/45	74.3 (10.2)	10.2 (3.6)	11.1 (5.1)	18.1 (5.7)
All (401)	169/232	68.0 (9.8)	10.9 (3.2)	18.4 (6.0)	24.5 (5.0)

At the time of preparing this paper, we have collected recordings from more than 500 subjects. Some of them were further filter out for the reasons like poor sound quality, interview interruption, heavy dialect, etc. Finally, 401 recordings from 401 subjects were obtained, including 138 healthy controls, 179 MCI patients and 84 dementia patients. The distributions of their age, education, MMSE and MoCA-BC scores are shown in Table I. This paper focuses on a binary classification between healthy control and dementia. Therefore, the data from MCI patients was not used in our study.

B. Task

This paper aims at building a dementia detector using spontaneous speech recorded in the picture description task. The picture description task was originally designed for the Boston Diagnostic Aphasia Examination [31]. It required each participant to say whatever happened in the picture (as shown in Fig. 1) as much as possible, and allowed encouragement from interviewer when participant had difficulties. The recording was conducted in a general clinic room with the door closed. For each subject, the interviewer and the subject's speech was recorded in a single audio file by a clip-on microphone placed

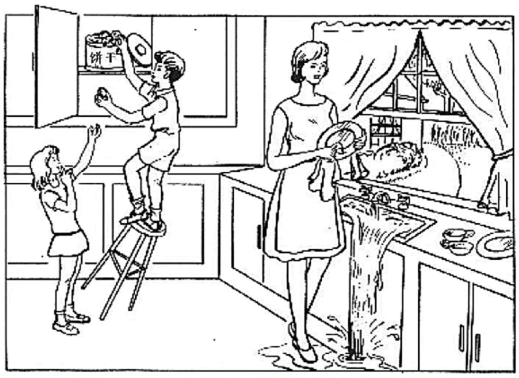

Fig. 1. The picture of "Cookie Theft", adopted from Boston Diagnostic Aphasia Examination [31]. The English word "cookie" is translated to Chinese word "饼干" during our data collection.

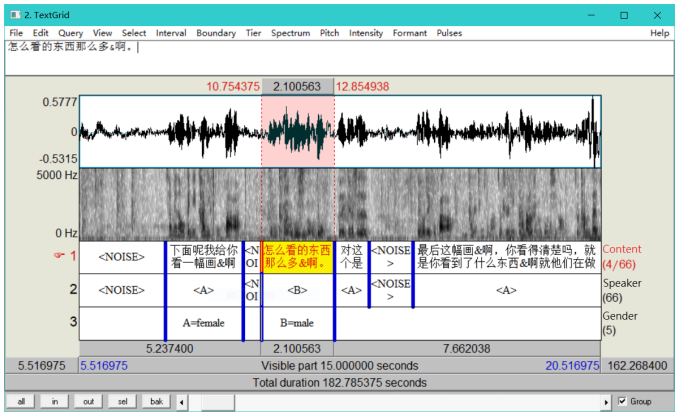

Fig. 2. The interface of annotation using the TextGrid format of Praat.

on the interviewer's collar in order to reduce the influences on subjects. The audio recordings were stored as 16-bit mono WAV format with a sampling rate of 16 kHz. The waveforms were then processed using a high-pass filter to filter out the low frequency noise below 60Hz.

C. Annotation

Speech samples were manually annotated using the TextGrid format of praat software [32], as shown in Fig. 2. The annotations included the transcription, the start and end time, and the speaker information of each sentence. Dialogues contained in the audio but not belonging to the picture description task were not transcribed. Unrecognizable sentences and non-speech segments such as laughter and cough were indicated using special tags.

Furthermore, we manually mark all occurrences of word repetition, word correction and grammatical errors. Filled pauses, usually indicating hesitation, were considered to be useful for dementia detection in previous study [11]. However, it is not so easy to distinguish the filled pause in Chinese. Therefore, we built a list of modal particles (including "噢", "哦", "啊", "嗯", "呃", "唉", "哎" in Chinese) and marked the occurrences of these words as filled pauses.

TABLE II
DESCRIPTIONS OF 16 DURATION FEATURES.

Description	Dimension
Total duration	1
Number of utterances	2
Duration of each utterance	2*2=4
Speak duration proportion	2
Silence duration proportion	1
Number of syllable	2
Articulation rate of each utterance	2*2=4
SUM	16

III. FEATURE EXTRACTION

Altogether 113 features of duration, acoustic, and linguistic categories are extracted from the speech waveforms together with their annotations to build our models for dementia detection. The details of them are introduced in this section.

A. Duration Features

Previous studies [17], [19], [20] have found that the dementia patients may have a low speech rate and frequent hesitations. Therefore, some features related with the durations of subjects and interviewers are extracted here using the utterance segmentations given by annotation. This category contains all together 16 features, as shown in Table II. Their descriptions are as follows.

- Total duration: The length of time from the start of the task to the end of the task.
- Duration of each utterance: We calculate the mean and standard deviation of utterance durations.
- Number of utterances: The total numbers of utterances spoken by the interviewer and by the participant.
- Number of syllables: The number of syllables in each utterance is estimated using Praat software (http://www.fon.hum.uva.nl/praat/) [33]. Then, their mean and variance are calculated.
- Articulation rate of each utterance (syllables per second): For each utterance, the number of syllables mentioned above is divided by the utterance duration. Then, we calculate their mean and variance.
- Speaking duration proportion: The duration of all utterances for the interviewer or from the participant divided by the total duration. If a participant has difficulties in this task, the interviewer's speech duration would be longer and the participant's speech duration would be shorter.
- Silence duration proportion: The duration of all silence segments divided by the total duration. It may indicate the hesitation of participant in this task.

B. Acoustic Features

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version (eGeMAPS) [34], which can be extracted using the open-source openSMILE

toolkit [35], have been widely used in speech emotion recognition and other tasks [36]. In this paper, the eGeMAPS features are adopted as our acoustic features.

This expert-knowledge-based set contains 88 features which is small and suited for small datasets. In the process of extracting acoustic features, we use overlapping windows, which are shifted forward at a rate of 10ms, to obtain 25 acoustic low level descriptors (LLDs) which cover common speech signal characteristics such as prosody (energy and F0), voice quality (jitter and shimmer) and Mel-frequency cepstral coefficients (MFCCs).

Before extracting acoustic features, we use annotations to identify and remove the audio segments that interviewer spoke. For each subject, we obtain an 88-dimensional vector through calculating the arithmetic mean and the coefficient of variation of LLDs over time. For the pitch, jitter, shimmer, and all formant related LLDs, only voiced regions are selected for calculation. Besides, some other statistical values, such as the 20-th, 50-th, and 80-th percentiles, the range of percentiles 20 – 80 and the mean and standard deviation of the slope of rising/falling signal parts, are calculated for F0 and loudness additionally.

C. Linguistic Features

1) Word counts and manually labelled features: Different form English, Chinese has no explicit word boundaries. Here, the open-source tool jieba [37] is used for word segmentation. After word segmentation, we count the number of word uttered by the interviewer and participants resulting in 2 features. Furthermore, the occurrences of filled pauses, word repetition, word correction and grammatical errors are also counted using the manual annotations described in Section II.

2) Perplexities derived from language models: Our previous study on the English DementiaBank dataset [38] have demonstrated that the perplexity features extracted by N-gram language models can benefit the automatic AD detection from continuous speech.

N-gram language models [39] have been widely used in the area of natural language processing. An N-gram model represents the conditional probability of

$$P\left(w_n|W_{n-1}^{n-N+1}\right) = P\left(w_n|w_{n-N+1},\ldots,w_{n-1}\right) \quad (1)$$

where $\{w_{n-N+1},\ldots,w_{n-1}\}$ are N-1 preceding words of word w_n in an utterance. The N-gram probabilities are estimated from a training corpus by counting the frequencies of words or word sequences. When an N-gram model λ is built, the perplexity value can be calculate to evaluate how likely a test sequence is generated by this model. A lower perplexity corresponds to a higher likelihood. For a test word sequence $X = \{w_1, w_2, \ldots, w_K\}$, its perplexity is defined as

$$\mathrm{PPL}(\lambda, X) = P(X|\lambda)^{-1/K}, \quad (2)$$

where

$$P(X|\lambda) = \prod_{n=1}^{K} P\left(w_n|W_{n-1}^{i-N+1},\lambda\right). \quad (3)$$

TABLE III
THE SUMMARY OF ALL FEATURES.

Feature category	Dimension
Duration	16
Acoustic	88
Linguistic	9
Demographic	4
SUM	117

It is expected that one text should achieve a low perplexity if it is evaluated by an N-gram language model (LM) trained using the training data of the same genre. Otherwise, the perplexity should be high if the training corpus and the test text are from different genres.

In our case of binary classification between Dementia and CTRL, the transcriptions of the control samples and the dementia samples in the training set are used to estimate two N-gram LMs λ_C and λ_D respectively. For a test sample with transcriptions X_i, a two-dimensional perplexity feature vector $\{PPL_C, PPL_D\}$ is calculated using λ_C and λ_D as

$$PPL_C = \mathrm{PPL}\left(\lambda_C, X_i\right), \quad (4)$$

$$PPL_D = \mathrm{PPL}\left(\lambda_D, X_i\right). \quad (5)$$

Furthermore, we calculate their difference as $PPL_{D-C} = PPL_D - PPL_C$ and form a three-dimensional feature vector $\{PPL_C, PPL_M, PPL_{D-C}\}$. Here, the unigram models of λ_C and λ_D are used according to the results of previous study [38].

D. Summary of features

The demographic attributes of subjects including age, gender and education are also used as features for dementia detection in our study. Note that one-hot encoding is adopted for the binary gender. This leads to 4-dimensional demographic features. Finally, there are 117 features altogether as shown in Table III.

IV. Classifier

Logistic regression (LR) is employed to build our classifiers for distinguishing dementia from control samples. The models are implemented using the Scikit-Learn toolkit [40]. The extracted features are standardized before classification assuming that all numerical features are centered around 0 and have unit variance. L1 or L2 penalty term is added to reduce the degree of overfitting and the penalty terms are defined as follows [41]

$$L1 : \|\boldsymbol{w}\|_1 = \sum_{j=1}^{m} |w_j|, \quad (6)$$

$$L2 : \|\boldsymbol{w}\|_2^2 = \sum_{j=1}^{m} w_j^2, \quad (7)$$

where m is the number of used features and w_j is the weight of the j-th dimension in the logistic regression model.

TABLE IV
PERFORMANCE OF LR CLASSIFIERS USING FEATURES OF DIFFERENT CATEGORIES.

Category	Penalty	Set	Accuracy	Precision	Recall	F1 score
Duration	l1	train	0.7822	0.7540	0.7993	0.7757
		test	0.7272	0.6947	0.7478	0.7046
Acoustic	l2	train	0.8243	0.8123	0.8323	0.8218
		test	0.6451	0.6433	0.6537	0.6294
Linguistic	l2	train	0.8083	0.7897	0.8203	0.8045
		test	0.7787	0.7577	0.7961	0.7601
Demographic	l1	train	0.6906	0.6937	0.6898	0.6913
		test	0.6685	0.6644	0.6683	0.6528
All	l1	train	0.8502	0.8548	0.8477	0.8509
		test	**0.8189**	**0.8187**	**0.8196**	**0.8086**

V. EXPERIMENTS

A. Experimental Conditions

Considering the imbalanced distribution between Dementia and CTRL samples in our dataset, a resampling strategy was adopted in our experiments. The dementia samples was kept all the time and we randomly selected 84 samples from the CTRL group. Then, experiments were conducted using the balanced dataset by 10-fold cross-validation (CV). We repeated the process of selecting CTRL samples using different random seeds by 20 times and reported the average results in order to reduce the fluctuations caused by resampling.

For extracting perplexity features, one practice concern is that one sample should not be used to train the language model which calculates the perplexity of itself. In each fold of CV, all perplexities of test samples were calculate using the LMs trained on the train set. For the samples in training set, we used 9-fold nested cross-validation, which meant that the perplexities of the samples belonging to one fold were calculated use the LMs trained on the other 8 folds. For building LR classifiers, the penalty factor C was chosen among $\{0.01, 0.1, 1.0, 10, 100\}$ by 4-fold cross validation in the train set for each fold.

B. Classification Performance

The accuracy, precision, recall and F1-score of the positive class (Dementia class) were adopted as metrics to evaluate the performance of built classifiers. The results of classification using feature of different categories are shown in Table IV. For each feature category, both penalty types (L1 or L2) were tried and the one with better overall accuracy is shown in the table.

It can be observed that the model using linguistic features obtained the best performance (test set accuracy of 77.9%) among all models using features of single category. The model using acoustic features performed even worse than the one using demographic features. More detailed analysis on the importance of different feature categories will be introduced in next subsection. Anyway, we can see that fusing the features of all categories achieved the best test set accuracy of 81.9%. This result is comparable with the performance achieved by previous studies on the datasets of other languages [12]–[15], [18], [38].

C. Feature Importance Analysis

Here, we analyzed the importance of different features using the metric of area under curve (AUC), which is defined as the area under the receiver operating characteristic (ROC) curve [31]. The value of AUC equals to the probability that when randomly choosing a pair of positive (dementia) and negative (healthy control) samples, the positive sample's feature value is larger than that of the negative sample. Thus if a feature has a good discriminating capability, its AUC should be close to 0 or 1. On the contrary, the AUC of a random feature should be around 0.5.

We first assumed that the feature and the label were positively correlated and calculated its AUC. If the result was less than 0.5, which meant the feature and the label were negatively correlated, we switched the class label of positive or negative and recalculate the AUC. The AUC values of different features are listed in Table V. Here, the features whose AUC values were below 0.6 are not shown. From this table, we can see that the PPL_{D-C} feature achieved the best discriminating ability for dementia detection, followed by age, the proportion of participant's speaking time, the average length of each sentence of the interviewer, and the total number of words spoken by the interviewer. Most of them are from the linguistic and duration categories.

We also collected the coefficients in the LR model using all features and averaged them across all folds and all repetitions. The features with the top-10 largest absolute coefficients are shown in Table VI. We can see that the coefficient vector is very sparse due to the L1 regularization. Actually, these top-10 features accounted for 87.2% of the total absolute coefficients for all features. Fig. 3 shows the proportions of the absolute coefficients corresponding to different feature categories in the LR model using all features. We can see that linguistic features accounted the largest proportion while the

TABLE V
AREA UNDER CURVE (AUC) VALUES OF DIFFERENT FEATURES. THE FEATURES WHOSE AUC VALUES WERE BELOW 0.6 ARE NOT SHOWN. THE INTERVIEWER AND THE SUBJECT ARE DENOTED AS A AND B RESPECTIVELY FOR ABBREVIATION.

Feature description	Category	AUC	Correlation
PPL_{D-C}	Linguistic	0.8476	+
Speaking duration proportion of B	Duration	0.7538	-
Number of words said by A	Linguistic	0.7451	+
Age	Demographic	0.7450	+
Number of syllables said by A	Duration	0.7247	+
Duration of A's utterances (mean)	Duration	0.7207	+
Total duration	Duration	0.7108	+
Duration of B's utterances (mean)	Duration	0.6927	-
Number of utterances said by A	Duration	0.6808	+
PPL_{D}	Linguistic	0.6791	+
Speaking duration proportion of A	Duration	0.6526	+
Education	Demographic	0.6430	-
mfcc1_sma3_stddevNorm	Acoustic	0.6197	-
Number of B's filledpauses	Linguistic	0.6194	+
Slience duration proportion	Duration	0.6164	+
Articulation rate of B's utterances (mean)	Duration	0.6147	-
Articulation rate of A's utterances (std)	Duration	0.6143	-
Duration of B's utterances (std)	Duration	0.6133	-
mfcc1V_sma3nz_stddevNorm	Acoustic	0.6120	-
Number of utterances said by B	Duration	0.6063	+
Articulation rate of B's utterances (std)	Duration	0.6062	+

TABLE VI
FEATURES WITH THE LARGEST ABSOLUTE COEFFICIENTS IN THE LR MODEL USING ALL FEATURES.

Feature description	Category	Coefficient
PPL_{D-C}	Linguistic	0.8285
Age	Demographic	0.4384
Speaking duration proportion of B	Duration	-0.2483
Number of words said by A	Linguistic	0.1631
Duration of A's utterances (mean)	Duration	0.1529
Number of grammatical errors made by B	Linguistic	0.0350
mfcc1V_sma3nz_stddevNorm	Acoustic	-0.0301
shimmerLocaldB_sma3nz_ stddevNorm	Acoustic	0.0279
F3bandwidth_sma3nz_amean	Acoustic	-0.0271
shimmerLocaldB_sma3nz_amean	Acoustic	-0.0161

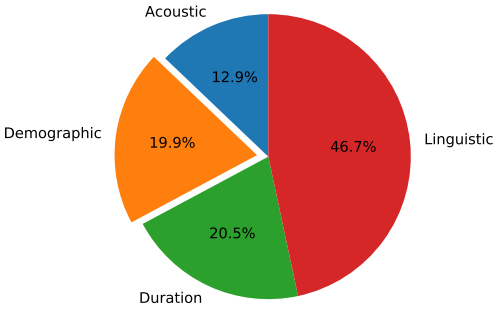

Fig. 3. Coefficient proportions of different feature categories.

proportion of acoustic features was the smallest. These results are consistent with the ones shown in Table IV.

D. Dementia Detection Using Cognitive Test Scores

Cognitive tests, such as MMSE and MoCA-BC, are common approaches for dementia screening nowadays. These tests need professional interviewers and are much more time-consuming than the picture description task. In our experiments, we also built an LR classifier using MMSE and MoCA-BC scores together with demographic attributes to examine its performance in comparison with our proposed method. The results are shown in Table VII. We can see that the test set accuracy of 91.4% was obtained which was about 10% higher than our best result in Table IV. To further refine our proposed method and make its result comparable with the performance of using cognitive test scores will be the goal of our future work.

TABLE VII
PERFORMANCE OF THE LR CLASSIFIER (PENALTY=L1) USING MMSE AND MOCA-BC SCORES TOGETHER WITH DEMOGRAPHIC FEATURES.

Set	Accuracy	Precision	Recall	F1 score
train	0.9308	0.8964	0.9629	0.9281
test	0.9137	0.8775	0.9465	0.9037

VI. CONCLUSIONS

This paper first introduces the Mandarin speech dataset we built for the study on dementia detection from spontaneous speech. Then, logistic regression classifiers are built using the speech features of different categories. Finally, the best test set accuracy of 81.9% is achieved in our experiments by using all features. Further analysis on the importance of different features reveals that linguistic features, especially the perplexity features, play the most important role in our model. To further increase the size of our dataset, to develop more sophisticated classifiers for dementia detection, and to replace manual annotation with automatic speech diarization and recognition will be the tasks of our future work.

ACKNOWLEDGMENT

This work was partially supported by the National Key R&D Program of China (No. 2018YFC1314700), the National Science Foundation of China (No. 81671307 and 61871358), and the Priority of Shanghai Key Discipline of Medicine (No. 2017ZZ02020).

REFERENCES

[1] W. H. Organization, "World health organization (2017) dementia fact sheet," 2017. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/dementia

[2] G. Waldemar, K. T. Phung, A. Burns, J. Georges, F. R. Hansen, S. Iliffe, C. Marking, M. O. Rikkert, J. Selmes, G. Stoppe *et al.*, "Access to diagnostic evaluation and treatment for dementia in europe," *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, vol. 22, no. 1, pp. 47–54, 2007.

[3] J. Jia, F. Wang, C. Wei, A. Zhou, X. Jia, F. Li, M. Tang, L. Chu, Y. Zhou, C. Zhou *et al.*, "The prevalence of dementia in urban and rural areas of china," *Alzheimer's & Dementia*, vol. 10, no. 1, pp. 1–9, 2014.

[4] J. Xu, J. Wang, A. Wimo, L. Fratiglioni, and C. Qiu, "The economic burden of dementia in china, 1990–2030: implications for health policy," *Bulletin of the World Health Organization*, vol. 95, no. 1, p. 18, 2017.

[5] R. G. Morris, *The cognitive neuropsychology of Alzheimer-type dementia.* Oxford University Press, 1996.

[6] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[7] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," *Cortex*, vol. 55, pp. 122–129, 2014.

[8] 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.

[9] V. Taler and N. A. Phillips, "Language performance in alzheimer's disease and mild cognitive impairment: a comparative review," *Journal of clinical and experimental neuropsychology*, vol. 30, no. 5, pp. 501–556, 2008.

[10] V. D. Santos, P. A. Thomann, T. Wüstenberg, U. Seidl, M. Essig, and J. Schröder, "Morphological cerebral correlates of cerad test performance in mild cognitive impairment and alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 23, no. 3, pp. 411–420, 2011.

[11] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

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

[13] T. Warnita, N. Inoue, and K. Shinoda, "Detecting Alzheimer's disease using gated convolutional neural network from audio data," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. September, pp. 1706–1710, 2018.

[14] S. Wankerl, E. Nöth, and S. Evert, "An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language." in *INTERSPEECH*, 2017, pp. 3162–3166.

[15] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of alzheimer's disease using neural network language models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5841–5845.

[16] J. Weiner, C. Frankenberg, D. Telaar, B. Wendelstein, J. Schröder, and T. Schultz, "Towards automatic transcription of ilse—an interdisciplinary longitudinal study of adult development and aging." in *LREC*, 2016.

[17] J. Weiner, M. Engelbart, and T. Schultz, "Manual and automatic transcriptions in dementia detection from speech." in *INTERSPEECH*, 2017, pp. 3117–3121.

[18] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in alzheimer's disease," *International journal of speech-language pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[19] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[20] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[21] M. F. Folstein, L. N. Robins, and J. E. Helzer, "The mini-mental state examination," *Archives of general psychiatry*, vol. 40, no. 7, pp. 812–812, 1983.

[22] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[23] J. C. Morris, "The clinical dementia rating (cdr): current version and scoring rules." *Neurology*, 1993.

[24] M. P. Lawton and E. M. Brody, "Assessment of older people: self-maintaining and instrumental activities of daily living," *The gerontologist*, vol. 9, no. 3_Part_1, pp. 179–186, 1969.

[25] R. H. Benedict, D. Schretlen, L. Groninger, and J. Brandt, "Hopkins verbal learning test–revised: Normative data and analysis of inter-form and test-retest reliability," *The Clinical Neuropsychologist*, vol. 12, no. 1, pp. 43–55, 1998.

[26] Q. Zhao, Q. Guo, F. Li, Y. Zhou, B. Wang, and Z. Hong, "The shape trail test: application of a new variant of the trail making test," *PLoS One*, vol. 8, no. 2, p. e57333, 2013.

[27] J. C. Borod, H. Goodglass, and E. Kaplan, "Normative data on the boston diagnostic aphasia examination, parietal lobe battery, and the boston naming test," *Journal of Clinical and Experimental Neuropsychology*, vol. 2, no. 3, pp. 209–215, 1980.

[28] J. E. Meyers and K. R. Meyers, "Rey complex figure test under four different administration procedures," *The Clinical Neuropsychologist*, vol. 9, no. 1, pp. 63–67, 1995.

[29] M. Hamilton, "The hamilton rating scale for depression," in *Assessment of depression*. Springer, 1986, pp. 143–152.

[30] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's as-

sociation workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263–269, 2011.
[31] H. Goodglass, E. Kaplan, and B. Barresi, *Boston Diagnostic Aphasia Examination Record Booklet*. Lippincott Williams & Wilkins, 2000.
[32] "Textgrid file formats," 2018. [Online]. Available: http://www.fon.hum.uva.nl/praat/manual/TextGrid_file_formats.html
[33] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.
[34] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
[35] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.
[36] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988258
[37] "Jieba - chinese text segmentation," 2018. [Online]. Available: https://github.com/fxsjy/jieba
[38] Z. Guo, Z. Ling, and Y. Li, "Detecting alzheimer's disease from continuous speech using language models," *Journal of Alzheimer's Disease, Accepted*, 2019.
[39] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing," *Upper Saddle River, NJ: Prentice Hall*, 2008.
[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[41] S. Raschka and V. Mirjalili, *Python machine learning*. Packt Publishing Ltd, 2017.