World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:17, No:11, 2023

Using Speech Emotion Recognition as a Longitudinal Biomarker for Alzheimer's Disease

Yishu Gong, Liangliang Yang, Jianyu Zhang, Zhengyu Chen, Sihong He, Xusheng Zhang, Wei Zhang

Abstract—Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects millions of people worldwide and is characterized by cognitive decline and behavioral changes. People living with Alzheimer's disease often find it hard to complete routine tasks. However, there are limited objective assessments that aim to quantify the difficulty of certain tasks for AD patients compared to non-AD people. In this study, we propose to use speech emotion recognition (SER), especially the frustration level as a potential biomarker for quantifying the difficulty patients experience when describing a picture. We build an SER model using data from the IEMOCAP dataset and apply the model to the DementiaBank data to detect the AD/non-AD group difference and perform longitudinal analysis to track the AD disease progression. Our results show that the frustration level detected from the SER model can possibly be used as a cost-effective tool for objective tracking of AD progression in addition to the Mini-Mental State Examination (MMSE) score.

Keywords—Alzheimer's disease, Speech Emotion Recognition, longitudinal biomarker, machine learning.

I. Introduction

ALZHEIMER'S disease is a progressive and irreversible neurological disorder that primarily affects the brain, causing a decline in memory, thinking abilities, and overall cognitive function. AD gradually damages and destroys brain cells, leading to significant cognitive impairment and behavioral changes. While the exact cause of Alzheimer's is not fully understood, it is believed to be influenced by a combination of genetic, environmental, and lifestyle factors. The early stages of AD are often characterized by mild memory loss, confusion, and difficulty with language and problem-solving. Persons living with AD experience changes in the brain's temporal lobe that affect their ability to process language. Patients with AD may show a decline in formal language abilities, including vocabulary, comprehension, and speech production. Patients may exhibit symptoms including word loss, inability to follow storylines, decreased speech, confusion in conversations, etc. In addition, as patients experience these changes in their abilities, the feeling of loss of control can build up frustration and anger [1], [2]. It is unsurprising that patients become frustrated at their loss of self-expression, and studies have demonstrated that impaired communication is strongly linked with the development of significant behavioral concerns [3].

Yishu Gong is with Harvard T.H. Chan School of Public Health, Boston, MA 02115 USA (e-mail: yishugong@hsph.harvard.edu).
Liangliang Yang is with Washington State University, Pullman, WA 99164 USA (e-mail: liangliang.yang@wsu.edu).
Jianyu Zhang is with University of Michigan, Ann Arbor, MI 48109 USA (e-mail: zjianyu@umich.edu).
Zhengyu Chen is with Zhejiang University, Hangzhou, Zhejiang, 310058 P. R. China (e-mail: chenzhengyu@zju.edu.cn).
Sihong He is with University of Connecticut, Storrs, CT 06269 USA (e-mail: sihong.he@uconn.edu).
Xusheng Zhang, is with Penn State University, State College, PA 16801 USA (e-mail: xushengz@psu.edu).
Wei Zhang, is with Harvard University, Cambridge, MA 02138 USA (e-mail: wei_zhang@g.harvard.edu).

Patients may seem overall positive, but subtle changes that indicate frustration may be difficult for humans to detect, especially in early AD. Therefore, we aim at exploring whether the frustration level detected from machine learning algorithms can be used as an indicator of patients' difficulty in completing a speech- and language-based task.

Automatic Speech Emotion Recognition (SER) has received much attention with features ranging from traditional features such as pitch, words per minute, etc to image-based features MFCC (Mel Frequency Cepstral Coefficients), Chroma, etc. The learning methods also include support vector machines, XGBoost, CNN, RNN, and so forth [4]. However, people may experience several different emotions at the same time or go through emotional changes if they are talking for a long period of time.

In this research paper, we propose to build an SER model using data from the IEMOCAP dataset [5]. We use Mel-frequency cepstral coefficients (MFCC) as features for SER, as they capture the spectral characteristics of speech signals. Then we apply our SER model to picture description recordings from AD and non-AD patients in DementiaBank Pitt Corpus [6]. Our SER model is capable of tracking emotional changes throughout a monologue instead of assigning one class to the monologue with high probability. Statistics are carried out to compare the different emotion percentages that AD and non-AD patients exhibit during their picture descriptions. Additionally, we explore if any of the emotion percentages change significantly with age in the AD and non-AD groups.

II. Methods

Researchers have employed various approaches in emotion classification using IEMOCAP. Poria et al. [7] utilized CNNs for audio and text data. Shor et al. [8] explored initial steps and challenges. Neumann et al. [9] aimed to enhance classification with novel techniques. Li et al. [10] introduced "EmoCaps", Hu et al. [11] presented "UNIMSE", and Kim et al. [12] proposed "EmoBERTa" fine-tuned on IEMOCAP. In addition, SpeechBrain [13] is an open-source toolkit for this purpose. These contributions advance emotion classification in IEMOCAP. Some also applied SER on DementiaBank to help with AD classification [14], [15]. However, most of these readily available models suffer from one or more of the following problems for emotion tracking:

1) lack of frustration as a class: We recognize ”frustrated” is an important emotion for understanding dementia [3]. Especially when patients perform picture description tasks, they may experience difficulties in recalling certain words, and they could end up sounding frustrated.
2) over-training: The phenomenon of over-training refers to focusing on assigning the audio clips to one of the classes with high confidence. This is not ideal for tracking the change of emotion over time since it results in fast switching between different emotions even in a short period of time.
3) recording length very different between the IEMOCAP emotion training dataset and the DementiaBank dataset: Each audio clip in the IEMOCAP dataset is only 3-5s long while the recordings in DementiaBank are usually more than 30s. It is possible the emotion is changing as the person finishes his/her narrative.
4) lack of privacy protection: Patients usually are more hesitant to provide actual recordings. Images converted from audio recordings are less identifiable.

Therefore we would like to build an emotion recognition model using transfer learning that is more suitable for understanding the percentage of emotions and captures smooth transitions over time. Furthermore, we would like to use extracted features from the audio to accomplish emotion classification and tracking without using the transcript or the audio itself for privacy protection.

A. Data Source

a) DementiaBank Pitt Corpus: This study specifically uses the Pitt Corpus, gathered longitudinally between 1983 and 1988 on a yearly basis as part of the Alzheimer Research Program at the University of Pittsburgh [16]. Participants were categorized into three groups: dementia, control (non-AD), and unknown status. All participants were required to be above 44 years of age, have at least seven years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial MMSE score of 10 or more, and be able to provide informed consent. This study selected only dementia and control groups for a binary diagnosis of AD and non-AD. In addition, we specifically chose the Cookie Theft description task subset. Participants were shown the Cookie Theft picture (Fig. 1) and were asked to describe the picture in their own words. Table I lists the data available in this set.

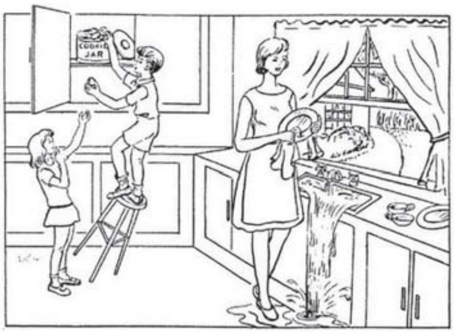


Fig. 1 The Cookie Theft picture from the Boston Diagnostic Aphasia Examination [17]

b) Emotions training data: In order to train machine learning classifiers to recognize emotions that include frustration, we employ the widely-used IEMOCAP dataset [5]. This dataset consists of professional actors expressing a variety of different emotions. This dataset contains 5 sessions. Each session contains scripted and improvised conversations between a different pair of male and female participants. In order to maximize our sample sizes during model training, we use both scripted and improvised audio clips. The recordings are clipped to only include one speaker at a time and then were then classified by listeners into 10 emotion categories (”angry”, ”happy”, ”disgusted”, ”fear”, ”frustrated”, ”excited”, ”neutral”, ”sad”, ”surprised”, and ”others”). When annotators cannot reach a consensus, the audio sample is marked as ”xxx”. If the annotators agree on a certain emotion that is not the 10 categories, the sample will be marked as ”other”. The audio clips of IEMOCAP vary in duration but are typically in the 3-5 second range. The total utterance is 100039 with the distribution of emotion shown in Fig. 2a.

We removed classes ”others” and ”xxx” since these two labels are not useful. We also discarded classes ”fear”, ”disgusted”, and ”surprised” due to small sample sizes. Lastly, similar to previous work on the IEMOCAP dataset [7, 9], we merged the ”excited” and ”happy” classes into a single class for which we use the ”happy” label. The resulting class has 7380 utterances with class breakdown shown in Fig. 2b.

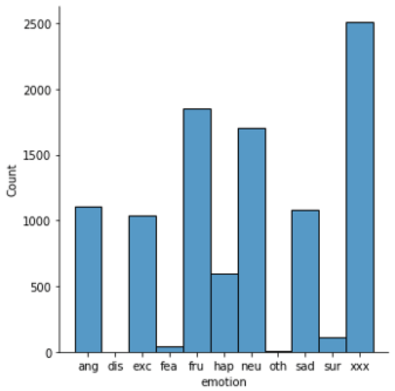
B. Data Preprocessing

For every audio recording, we trimmed out the interviewers' voices. Then, audio files were transformed to an MFCC (Mel Frequency Cepstral Coefficients) representation for use in machine learning, using the librosa library [18].

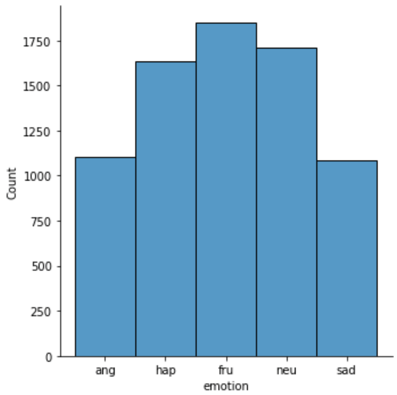
MFCC values were plotted to form images, which were then normalized and resized to meet the expected image size for machine learning classifiers. MFCC images had a fixed

TABLE I
STATISTICS OF PARTICIPANTS FROM DEMENTIABANK PITT CORPUS

PATTYPE	gender	age (first visit)	MMEScore (first visit)	N (patients)	2 recordings	3 recordings	4 recordings	5 recordings
AD	female	72.0 ± 9.0	19.6 ± 4.9	126	49	11	4	0
	male	69.7 ± 8.1	20.9 ± 5.7	66	26	13	8	3
non-AD	female	63.2 ± 8.9	29.2 ± 1.1	58	46	31	16	2
	male	64.4 ± 8.3	28.9 ± 1.2	41	29	15	2	2



(a)



(b)

Fig. 2 (a) Distribution of all emotion classes from the IEMOCAP dataset before filtering; (b) Distribution of all emotion classes from the IEMOCAP dataset after filtering

duration of 5 seconds, chosen because the typical length of IEMOCAP utterances is close to 5 sec. Shorter utterances were zero-padded to 5 sec, while longer utterances were truncated.

C. Emotions Recognition Pre-Training and Statistical Analysis

After we converted these utterances into MFCCs, standard transfer learning techniques were applied to MFCCs with two base models: VGG11 [19] and AlexNet [20]. During the training, we kept session 5 as the test set to assess performance and used sessions 1, 2, and 3 as the training set. Hyper-parameter tuning and model ensemble selection were performed using session 4 as a validation set. During the process of training, we further addressed the slight class imbalance with Synthetic Minority Oversampling Technique (SMOTE) [21].

In order to prevent the over-training phenomenon, our maximum training epoch is set to 20, which is much smaller compared to other similar works since these two are considered major risk factors for dementia [22]. Then early stopping is used to obtain the optimal epoch number.

We apply the emotion classifier on each recording to track emotion progression. The analysis pipeline can be summarized in a flowchart given in Fig. 3.

We use propensity score to match the AD patients with healthy participants with age and gender [23]. For baseline analysis, the mean emotion percentages across each utterance were computed. A T-test with Bonferroni correction is applied to determine statistical significance. F-test is computed for variances of mean emotion over time. We also conducted a subgroup analysis to study gender differences.

For longitudinal analysis, we only include patients who have completed at least two visits and one-to-one matched the AD with the non-AD group using propensity score with gender and age at the first visit. We use a linear mixed-effect model to perform a longitudinal analysis of frustration percentages. In (1), Gender = 0 refers to 'female' and Gender = 1 refers to 'male; Type = 0 refers to 'AD' and Type = 1 refers to 'non-AD'.

$$Fru_{id,age} = \alpha_1 \cdot Gender_{id} + \alpha_2 \cdot Type_{id} + \alpha_3 \cdot Age_{id,age} + \alpha_4 \cdot Type_{id} \times Age_{id,age} + \beta + \underbrace{\alpha_{id} \cdot Age_{id,age} + \beta_{id}}_{\text{random effects}} \quad (1)$$

Fixed-effects include an intercept, gender, age, and type, as well as interaction between age and type. Random-effects include age and intercept. Note that we normalized the age factor in the linear mixed effect model.

III. Results

A. Emotion Classification

Using the two base models AlexNet and VGG11, we ensemble them using average probability and we refer to it as the Ensemble Model in the following text. On test set session 5, our Ensemble Model classifies 5 classes (angry, happy, sad, neutral, frustrated) with an accuracy of 45% and a top 2 classes accuracy of 72%.

The confusion matrix is shown in Fig. 4. As we can see on the confusion matrix, the model is experiencing difficulty differentiating neutral from other emotions. We also present the Receiver operating characteristic for each class in Fig. 5. The micro-averaged AUC is 0.76.

For comparison, the commercially developed wav2vec2-IEMOCAP model is capable of classifying 4 classes (angry, happy, sad, neutral) with an accuracy of 75% [13]. We would like to emphasize that the purpose of our Ensemble Model is to understand emotion percentage and track emotion transition smoothly over time instead of assigning audio to a single class with high confidence. This usage will be elucidated when applying our Ensemble Model to long speech recordings ($\geq$ 30s) in the next section.

B. Emotion Progression Tracking

For every sliding window of 5s we have emotion percentage for 5 classes: happy (hap), neutral (neu), frustrated (fru), angry (ang), and sad (sad). We can visualize the emotion progression as the participants perform the picture description task. In Fig. 6, we present the emotion tracking result for a group of age and gender-matched AD and non-AD patients as examples at their first visit. As we can see in Fig. 6a, the AD patient became more frustrated (purple section) and angry (red section) in the

World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:17, No:11, 2023

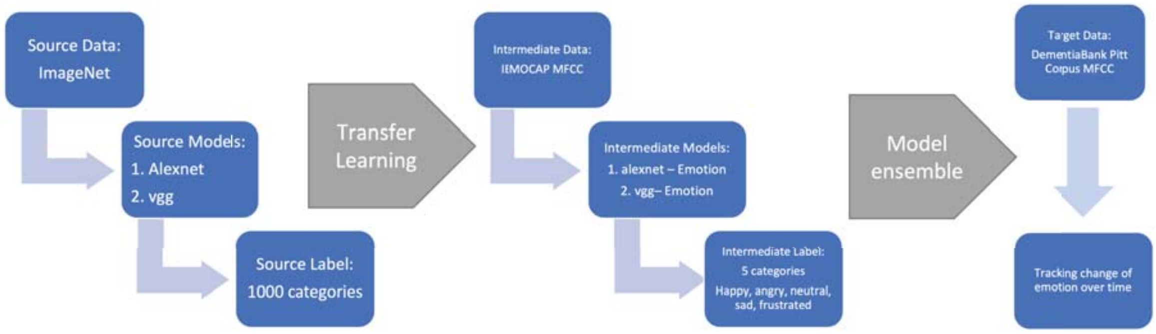


Fig. 3 Flowchart of Transfer Learning for emotion tracking

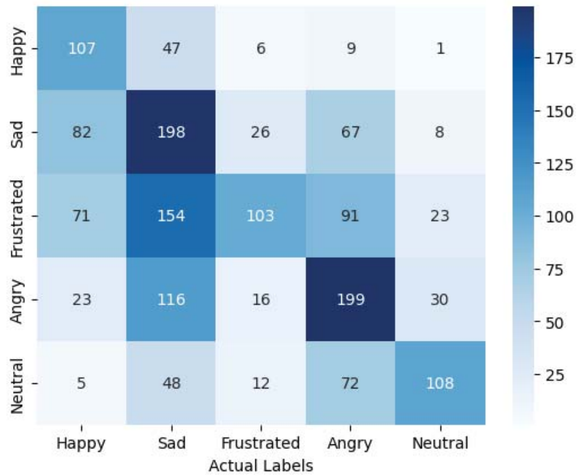


Fig. 4 Confusion matrix of Ensemble Model on the test set

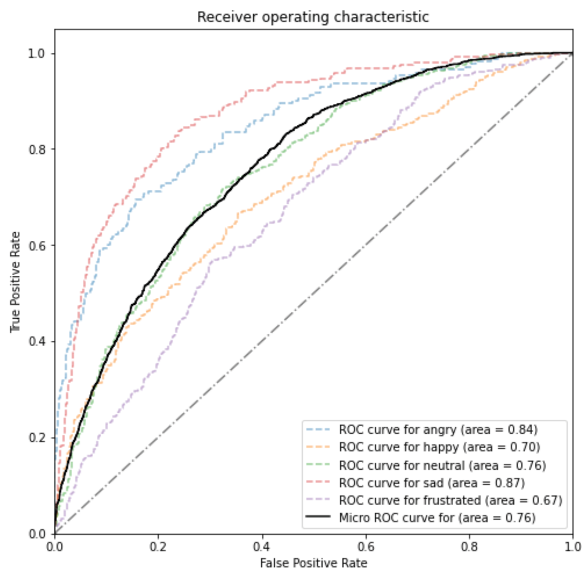
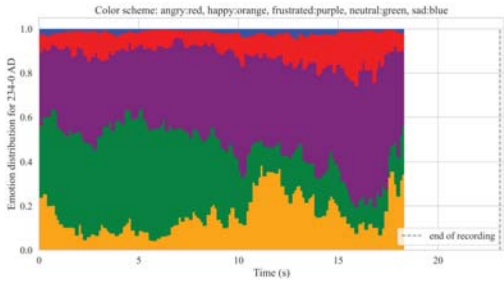
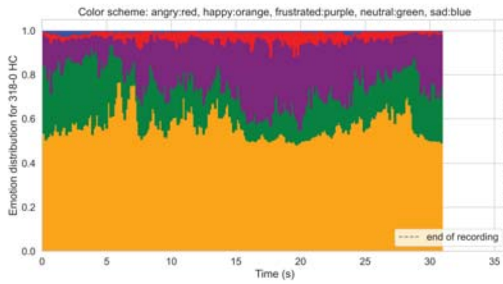


Fig. 5 ROC of Ensemble Model on the test set



(a)



(b)

Fig. 6 (a) Example emotion progress of an AD patient when completing the cookie theft description task; (b) Example emotion progress of a non-AD patient when completing the cookie theft description task (Note that these two patients are gender and age-matched)

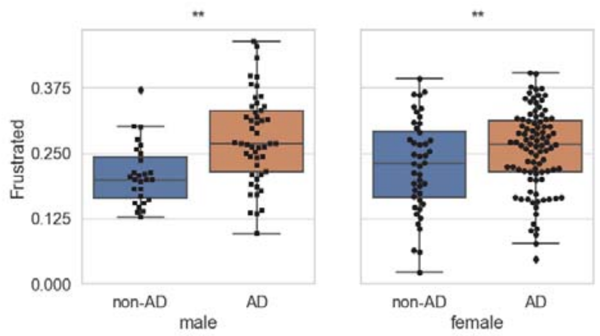
later half of the recording while in Fig. 6b the non-AD patient is mostly happy (orange section) when completing the picture description task. From the emotion tracking result, we can compute the average percentage of each emotion class over time.

a) Baseline Comparison: From Table II, we see that for the picture description task, the AD group has a statistically significant higher average frustrated and angry percentage over time than the non-AD group in both genders as shown in Fig. 7 (T-test was performed for all emotions, then p-values were corrected using the Bonferroni method). In addition, the F-test two sample for variances test shows frustrated percentage has a significantly smaller variance compared to that of happy and neutral percentages (comparing frustrated and happy: $p < 0.001$, comparing frustrated and neutral: $p < 0.001$).

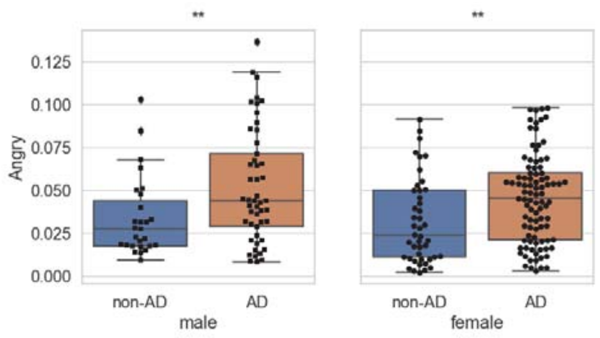
As research [2], [3], [24] has pointed out, it can be very frustrating for the person with AD or another dementia as they experience changes in their abilities. Feelings of loss of control and building frustration combined with the physical changes caused by the disease may cause the person to have emotional reactions to situations.

TABLE II
MEAN WITH A STANDARD DEVIATION OF AVERAGE EMOTION PERCENTAGE OVER TIME FOR MATCHED PARTICIPANTS AT THE FIRST VISIT PERFORMING COOKIE DESCRIPTION TASK GROUPED BY HEALTH STATUS AND GENDER

Gender	health status	N	angry	happy	sad	neutral	frustrated	recording length (s)
AD	Female	98	$4.5\% \pm 2.6\%$	$36.9\% \pm 20.0\%$	$1.3\% \pm 0.8\%$	$31.6\% \pm 15.6\%$	$25.6\% \pm 7.0\%$	31.6±16.5
	Male	54	$5.2\% \pm 3.3\%$	$36.4\% \pm 19.9\%$	$1.4\% \pm 0.8\%$	$29.2\% \pm 13.6\%$	$27.7\% \pm 8.5\%$	43.2±36.8
non-AD	Female	49	$3.2\% \pm 2.5\%$	$46.7\% \pm 17.9\%$	$1.1\% \pm 1.1\%$	$26.3\% \pm 14.9\%$	$22.7\% \pm 8.9\%$	37.1±23.1
	Male	27	$3.3\% \pm 2.3\%$	$45.0\% \pm 19.6\%$	$1.4\% \pm 0.8\%$	$29.3\% \pm 15.6\%$	$20.1\% \pm 5.8\%$	36.0±19.2



(a)



(b)

Fig. 7 Mean frustrated (a) and angry (b) emotion percentage over time for matched participants performing picture description task at the first visit grouped by type and gender. '**' means Bonferroni corrected p-values < 0.005, '*' means Bonferroni corrected p-values < 0.05, and 'N.S' means not statistically significant

When performing subgroup analysis, we observe the difference in frustrated and angry percentages in males is more pronounced than in females as shown in Fig. 7. Moreover, men with AD tend to have longer recording length (after trimming out interviewers' speech) compared to men in the non-AD group on average. However, the variance is too large to conclude statistical significance.

b) Longitudinal analysis: We present the fit result in Table III. As we can see, on average at age 66, AD patients have a 24.7% of frustration level when completing the cookie theft description while non-AD patients are 1% lower than AD patients. This observation coincides with what we observe in the baseline comparison. More importantly, as age progresses, AD patients' frustration level increases with age at the rate of 2.1% per year ($p < 0.01$) while no statistically significant change was found in the non-AD group.

We can visualize the linear mixed effect model in Fig. 8. Raw data and simple linear regression are presented on the top row and the predicted value from the linear mixed effect model is shown in the bottom row.

TABLE III
LINEAR MIXED EFFECT MODELING RESULT OF THE FRUSTRATION PERCENTAGE FOR PICTURE DESCRIPTION TASK COMPLETED AT DIFFERENT AGES (AGE IS NORMALIZED IN THE MODEL)

	Coef.	Std.Err.	z	$P > \|z\|$	[0.025	0.975]
β	0.247	0.008	29.180	0.000	0.230	0.263
α_1	-0.010	0.011	-0.925	0.177	-0.031	0.011
α_2	0.005	0.010	0.498	0.309	-0.014	0.024
α_3	0.021	0.008	2.512	0.006	0.005	0.037
α_4	-0.016	0.011	-1.451	0.074	-0.039	0.006

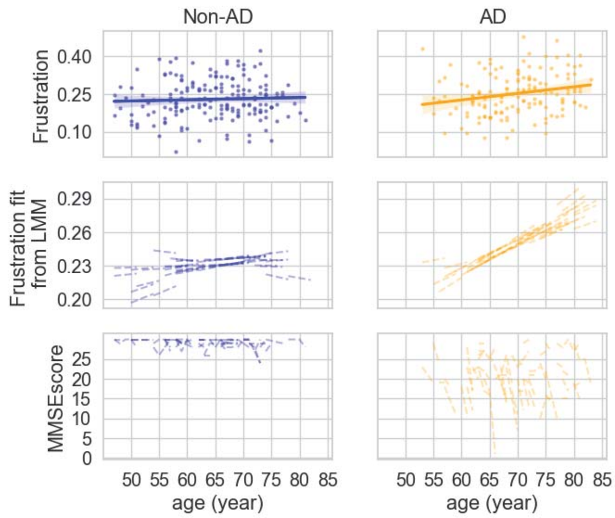


Fig. 8 Modeling frustration level when completing picture description task at different ages. Top row: raw data and simple linear regression model. Middle row: predicted individual progression from linear mixed effect modeling (Every dashed line indicates one patient). Bottom row: MMSE score (Every dashed line indicates one patient)

This faster increase in frustration percentage when describing the picture may be considered an indicator of the increased difficulty of completing the task as patients age. Indeed, when comparing MMSE scores between the two group, there is no significant change in the non-AD group while almost all patients in the AD group has a worse score as they age.

IV. DISCUSSION

We performed emotion tracking for picture description recordings in the Dementiabank Pitt corpus and found that the overall dominant emotions are happy and neutral. This is expected as there should not be any obvious negative perception in the Cookie Theft picture. For the non-AD group, the happy level is statistically higher than that of the AD group.

This indicates that non-AD patients tend to display a more positive attitude during testing. Nevertheless, the selection of description images could affect the overall distribution of different emotions. Only using the Cookie Theft picture could not suggest the absolute proportion of emotions within each subgroup. However, in this study, we focused on identifying the relative difference in frustration levels between AD and non-AD groups. Given all the frustration levels exceed 20% and rank 3rd in each subgroup shown in Table II, the statistical analysis and conclusion are trustworthy. Notably, the variance of frustration is way smaller than that of happy and neutral ($p < 0.001$ for both comparisons). It implies that when frustration shows up, the trend would be more persistent than other emotions. The stable pattern with longer time exhibition helps the therapist to capture the patient's frustration which may suggest a higher possibility of AD.

Among all emotions, it is more challenging to identify frustration, shown by the ROC curve in Fig. 5. Compared to happy or sad, which hold apparent positiveness or negativeness, the tone of frustration is more complex. Some patients could display frustration in a more dispirited way with a lowering mood. While others would show a more angry-like frustration, through a more irritable voice. Even in daily life, humans or doctors may have trouble distinguishing or defining frustration. In our study, we try to avoid classifying frustration precisely, instead, aim at quantifying its percentage out of other primary emotions. This treatment enables us to withstand the ambiguity of frustration, making our observation more applicable to various types of frustration.

It is commonly known that the risk of Alzheimer's increases with aging. Coincidentally, in this research, we identified older people in the AD group have a higher level of frustration. With the perception difficulty along with communication trouble becoming more serious as patients age and Alzheimer's disease develop, the emotional health would be affected negatively. When performing a task, the feeling of inability can be expressed by negative emotions, such as frustration. Such a phenomenon is evidenced by our daily observation that older Alzheimer patients are more vulnerable to emotional swings and hard to control them. At an individual patient level, our model predicts a faster increase in frustration level when age, which coincides with a faster increase of MMSE score measured. This raises a dangerous alert that the development of frustration could accelerate. It is worthwhile to study if the single patient expresses a growing degree of frustration at different phases of Alzheimer's disease, given relevant data sources available.

References

[1] Allan M Landes, Susan D Sperry, and Milton E Strauss. Prevalence of apathy, dysphoria, and depression in relation to dementia severity in alzheimer's disease. *The Journal of neuropsychiatry and clinical neurosciences*, 17(3):342–349, 2005.

[2] Fariba Mirakhori, Mina Moafi, Maryam Milanifard, Hossein Tahernia, et al. Diagnosis and treatment methods in alzheimer's patients based on modern techniques: The orginal article. *Journal of Pharmaceutical Negative Results*, pages 1889–1907, 2022.

[3] Michael Woodward. Aspects of communication in alzheimer's disease: clinical features and treatment options. *International psychogeriatrics*, 25(6):877–885, 2013.

[4] Inês Vigo, Luis Coelho, and Sara Reis. Speech-and language-based classification of alzheimer's disease: A systematic review. *Bioengineering*, 9(1):27, 2022.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[6] Francois Boller and James Becker. Dementiabank database guide. *University of Pittsburgh*, 2005.

[7] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[8] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*, 2020.

[9] Michael Neumann and Ngoc Thang Vu. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394. IEEE, 2019.

[10] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*, 2022.

[11] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*, 2022.

[12] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021.

[13] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[14] Fasih Haider, Sofia de la Fuente, Pierre Albert, and Saturnino Luz. Affective speech for alzheimer's dementia recognition. *LREC: Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID)*, pages 67–73, 2020.

[15] M Rupesh Kumar, Susmitha Vekkot, S Lalitha, Deepa Gupta, Varasiddhi Jayasuryaa Govindraj, Kamran Shaukat, Yousef Ajami Alotaibi, and Mohammed Zakariah. Dementia detection from speech using machine learning and deep learning architectures. *Sensors*, 22(23):9311, 2022.

[16] Jody Corey-Bloom and Michael S Rafii. The natural history of alzheimer's disease. In *Dementia*, pages 473–489. CRC Press, 2017.

[17] Louise Cummings. Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, 10(2):153–176, 2019.

[18] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[21] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[22] Juergen Dukart, Matthias L Schroeter, Karsten Mueller, and Alzheimer's Disease Neuroimaging Initiative. Age correction in dementia–matching to a healthy brain. *PloS one*, 6(7):e22193, 2011.

[23] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

[24] Madeline M Maier-Lorentz. Effective nursing intervention for the management of alzheimer's disease. *Journal of Neuroscience nursing*, 32(3):153, 2000.