Using Automatic Speech Recognition to Assess Thai Speech Language Fluency in Montreal Cognitive Assessment (MoCA)

Mr. Pimarn Kantithammakorn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2020
Copyright of Chulalongkorn University

การใช้เทคโนโลยีการรู้จำเสียงพูดแบบอัตโนมัติช่วยประเมินความสามารถทางภาษาของเสียงภาษาไทยจากแบบประเมินพุทธิปัญญาโมคา

นายพิมาน ขันติธรรมากร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	Using Automatic Speech Recognition to Assess Thai Speech Language Fluency in Montreal Cognitive Assessment (MoCA)
By	Mr. Pimarn Kantithammakorn
Field of Study	Computer Science
Thesis Advisor	Associate Professor Dr. PROADPRAN PUNYABUKKANA
Thesis Co Advisor	Assistant Professor Dr. DITTAYA WANVARIE

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

-- Dean of the FACULTY OF ENGINEERING

(Professor Dr. SUPOT TEACHAVORASINSKUN)

THESIS COMMITTEE

-- Chairman

(Dr. EKAPOL CHUANGSUWANICH)

-- Thesis Advisor

(Associate Professor Dr. PROADPRAN PUNYABUKKANA)

-- Thesis Co-Advisor

(Assistant Professor Dr. DITTAYA WANVARIE)

-- Examiner

(Associate Professor Dr. SOLAPHAT HEMRUNGROJN)

-- Examiner

(Dr. Chaipat Chunharas)

-- External Examiner

(Dr. Theerawit Wilaiprasitporn)

พิมาน ขันติธรรมากร : การใช้เทคโนโลยีการรู้จำเสียงพูดแบบอัตโนมัติช่วยประเมินความสามารถทางภาษาของเสียงภาษาไทยจากแบบประเมินพุทธิปัญญาโมคา. (Using Automatic Speech Recognition to Assess Thai Speech Language Fluency in Montreal Cognitive Assessment (MoCA)) อ.ที่ปรึกษาหลัก : โปรดปราน บุณยพุกกณะ, อ.ที่ปรึกษาร่วม : ฑิตยา หวานวารี

Montreal Cognitive Assessment (MoCA) เป็นแบบประเมินที่ได้รับการยอมรับอย่างแพร่หลายในการคัดกรองคนไข้ที่มีภาวะรู้คิดบกพร่องเล็กน้อยรวมถึงการประเมินความสามารถทางภาษาและการพูดโดยให้คนไข้พูดคำตามเงื่อนไขให้ได้มากที่สุดภายในระยะเวลาที่กำหนด โดยการคิดคะแนนจะนับคำที่ถูกต้องตามเงื่อนไขและไม่ซ้ำคำเดิมซึ่งอาจแตกต่างกันในแต่ละภาษา งานวิจัยชิ้นนี้ศึกษาการประเมินแบบทดสอบด้วยภาษาไทยโดยนำเทคนิคด้านการรู้จำเสียงพูดแบบอัตโนมัติมาช่วยในการคิดคะแนนของความสามารถทางภาษาในการทดสอบแบบประเมิน MoCA. ภาษาไทยเป็นภาษาที่มีข้อมูลเสียงที่สามารถนำมาใช้ได้แบบสาธารณะได้ค่อนข้างจำกัด โดยเฉพาะข้อมูลเสียงของคนไข้ที่มีภาวะรู้คิดบกพร่องเล็กน้อย เราจึงนำเสนอวิธีการสร้างแบบจำลองทางอะคูสติกด้วย Time Delay Neural Network - Hidden Markov Model (TDNN-HMM) มาช่วยในการพัฒนาระบบการรู้จำเสียงพูดแบบอัตโนมัติ ที่สามารถนำไปใช้ในสภาวะที่อาจมีเสียงรบกวนและคุณภาพเสียงของคนไข้อาจไม่ดีเท่าที่ควร โดยการนำข้อมูลเสียงภาษาไทยสาธารณะที่ชื่อว่า LOTUS มาช่วยในการพัฒนาโมเดลรวมทั้งขั้นตอนในการลดสัญญาณรบกวนออกจากไฟล์เสียงก่อนนำมาประมวณผลเพื่อไปใช้ในการนับคำและให้คะแนนในส่วนการประเมินความสามารถทางภาษาต่อไป ผลการทดลองแสดงให้เห็นว่า โมเดลแบบ TDNN-HMM ร่วมกับการเพิ่มปริมาณข้อมูลเสียง มาช่วยในการเรียนรู้คุณลักษณะแบบ lattice-free maximum mutual information (LF-MMI) ช่วยลดความผิดพลาดของคำที่ทำนายได้ โดยมีอัตราการผิดพลาดของคำอยู่ที่ประมาณ 41.30% ซึ่งยังไม่เคยมีงานวิจัยชิ้นใดเคยทำมาก่อนในการนำเทคนิคด้านการรู้จำเสียงพูดอัตโนมัติมาช่วยในการคิดคะแนนความสามารถทางภาษาสำหรับภาษาไทย

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต ....

ปีการศึกษา 2563 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

6270194021 : MAJOR COMPUTER SCIENCE

KEYWORD: MoCA ASR speech recognition scoring language fluency test

Pimarn Kantithammakorn : Using Automatic Speech Recognition to Assess Thai Speech Language Fluency in Montreal Cognitive Assessment (MoCA). Advisor: Assoc. Prof. Dr. PROADPRAN PUNYABUKKANA Co-advisor: Asst. Prof. Dr. DITTAYA WANVARIE

The Montreal Cognitive Assessment (MoCA), a widely accepted screening tool for identifying patients with mild cognitive impairment (MCI), includes a language fluency test of verbal functioning where scores are based on the number of unique correct words produced by the test-taker. However, with different languages, it is possible that unique words may be counted differently. This study focuses on Thai as a language that differs from English in its type of word combination. We applied various automatic speech recognition (ASR) techniques to develop an assisted scoring system for the language fluency test of the MoCA with Thai language support. The extra challenge is that Thai is a low-resource language where domain-specific data are not publicly available, especially speech data from patients with MCI. We propose a hybrid Time Delay Neural Network - Hidden Markov Model (TDNN-HMM) architecture for acoustic model training to create our ASR system that is robust to environmental noise and the variation of voice quality impacted by MCI. The LOTUS Thai speech corpus is incorporated into the training set to improve the model's generalization. A preprocessing algorithm is implemented to reduce the background noise and improve the overall data quality before feeding into the TDNN-HMM system for automatic word detection and language fluency score calculation. The results show that the TDNN-HMM model in combination with data augmentation using lattice-free maximum mutual information (LF-MMI) objective function provides a word error rate (WER) of 41.30%. To our knowledge, this is the first study to develop an ASR with Thai language support to automate the scoring system of the MoCA's language fluency assessment.

Field of Study:	Computer Science	Student's Signature ....
Academic Year:	2020	Advisor's Signature ....
		Co-advisor's Signature ....

ACKNOWLEDGEMENTS

My deepest gratitude goes to my advisor, Assoc. Prof. Proadpran Punyabukkana, co-advisor, Asst. Prof. Dittaya Wanvarie and Dr. Naruemon Pratanwanich for their guidance and support throughout this thesis.

I would like to thank Assoc. Prof. Dr. Solaphat Hemrungrojn, Dr. Chaipat Chunharas from the Faculty of Medicine, Chulalongkorn University; and the staffs at King Chulalongkorn Memorial Hospital on their relentless support and valuable insights on MoCA assessment and its evaluation process. Very special thanks to all of the Digital MoCA project team members who have been working very hard to develop a new exciting application. We have learned a great deal from this experience together. Last but not least, thanks to my family who are always there to support me through this challenge.

Pimarn Kantithammakorn

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

Introduction

Cognitive decline is a common health issue among the aging population [1]. Memory loss and forgetfulness are part of the normal aging process, but memory loss that affects daily life can be a symptom of dementia [2]. Neurodegeneration is one common cause of dementia [3], and symptoms vary among individuals, presenting as memory loss, deterioration of speech, motor skills, or cognitive function [4]. Mild cognitive impairment (MCI) is a condition in between normal age-associated memory impairment and dementia [5]. MCI causes cognitive problems that are noticeable by individuals and those close to them but do not impact daily life activities. As it is believed to be a high-risk condition for the development of Alzheimer's disease (AD) [6], early detection of MCI will allow health professionals to better plan for and treat individuals at risk of developing AD or other types of dementia.

A number of screening tools have been developed for detecting dementia. Among them, the Mini-Mental State Examination (MMSE) [7] is the most widely used, but difficulties in detecting early dementia have been reported. To address this problem, the Montreal Cognitive Assessment (MoCA) was developed to screen patients with MCI while performing on a normal range of MMSE [8]. In comparison, the MoCA showed a sensitivity of 90% where MMSE had a sensitivity of 18% in identifying MCI patients with memory loss [9]. The MoCA assesses several parts of brain function such as short-term memory, visuospatial abilities, executive functions, attention, concentration and working memory, language, and orientation. It is available in both paper and digital format, but data collection and scoring require extensive assistance by health professionals, and further data analysis is limited to text input only.

The "Digital MoCA" is a newly developed application for iPads based on the standard MoCA test criteria with Thai language support including automatic data collection and scoring with limited assistance from health professionals. Automatic data collection of voice recording and drawing input enable further data analysis using machine learning techniques for better understanding of MCI risk factors.

To achieve reliable results from the Digital MoCA, automated scoring is preferred to limit impacts from human error, personal judgment, and bias. Automatic speech recognition (ASR) can be adopted to translate voice recordings into sequences of words for automated scoring and evaluation. This research focuses on applying ASR techniques to detect proper Thai words to assist the automated scoring system of the language fluency test based on the MoCA test criteria and to reduce the need for health professionals to provide personal judgments during assessments, which will help to increase the overall reliability of MoCA scores. Developing ASR for intended functionalities in this situation comes with several challenges such as a high variety of pronunciation, tone, and accent of the patients, all of which might differ from healthy controls. In addition, background noises and conversations between patients and health professionals are recorded during the test. To overcome these challenges, several techniques need to be applied in combination with a special algorithm to differentiate between words that eligible for MoCA scoring versus those that are ineligible.

Development of the Digital MoCA

The MoCA is available in two versions, a standard paper-and-pencil version and an electronic version. Using the electronic MoCA enables opportunities for utilizing technology like artificial intelligence (AI) to significantly improve the efficiency and quality of cognitive assessment. To address the limitations of the standard MoCA test procedure for Thai, researchers from the Faculty of Medicine in collaboration with Department of Computer Engineering at Chulalongkorn University initiated a new project to develop a customized version of the MoCA as an application for iPads called the Digital MoCA. The main goal was to enable automatic data collection and utilize machine learning techniques for the early detection of MCI patients in Thailand. As of the time of this writing, the Digital MoCA is still in the early phase of development with a prototype version under evaluated by a group of physicians and a health professional team. The Digital MoCA application offers the same procedure and test criteria as the standard paper-and-pencil version but with an additional function for voice recording and image data capturing in each test module to store in the system database for further analysis with machine learning. This research paper is a sub-project of the Digital MoCA project initiative to utilize new AI techniques for MCI detection improvement.

Related Work

Many studies have applied speech processing to cognitive assessment. The common approach used for analysis relies on acoustic features from speech data and text features extracted through ASR. König et al. [10] analyzed dementia-related characteristics from voice and speech patterns by developing a classifier using a support vector machine (SVM) with features extracted from spoken tasks characterized by the continuity of speech using the duration of contiguous voice and silent segments and the length of contiguous periodic and aperiodic segments to derive statistical values as vocal features. For the semantic fluency task, the vocal feature was defined as the distance in time from the first detected word to each following word. Evaluation results showed the following classification accuracy: HC versus MCI, 79%; HC versus AD, 87%; and MCI versus AD, 80%. These findings suggested that automatic speech analyses could be an additional assessment tool for elderly patients with cognitive decline.

Spontaneous speech can provide valuable information about the cognitive state of individuals [11], but to retrieve clinical useful information, it needs to be transcribed manually. Zhou et al. [11] proposed an ASR system to generate transcripts automatically and extract text features to identify AD with an SVM classifier. They used an open-source Kaldi ASR toolkit [12] to optimize performance for speakers with and without dementia over the Dementia Bank (DB) corpus by insertion penalty and language model weight adjustment. They obtained an average WER of 38.24% that shown improvement over their previous work, which employed commercial ASR. However, their results were limited by the poor quality of audio in the DB corpus, necessitating further exploration.

Language fluency tests are one of the main tasks on cognitive tests; they are also known as verbal fluency (VF) tasks in the literature, referring to short tests of verbal functioning where patients are given one minute to produce as many unique words as possible within a sematic category (semantic fluency) or starting with a given letter (phonemic fluency) [13]. In clinical practice, VF tests are administered manually, and few studies have evaluated computerized VF administration and scoring. Pakhomov et al. [14] applied ASR to speech data collected during VF tasks to obtain an approximate count of legitimate words. They implemented an ASR system based on Kaldi [12] using a speaker-independent acoustic model with a specially trained animal fluency language model and applied confidence scoring to post-process ASR output. Standard manual scoring was performed, including the transcription of all responses during the VF task to be used for the evaluation of the ASR decoder performance. They achieved a WER of 56%, which was relatively high, but the results suggested that the combination of speaker adaptation and confidence scoring improved overall accuracy and was able to produce a VF estimated score that was very close to ones yielded by human assessment.

Tröger et al. [15] proposed telephone-based dementia screening with automated semantic verbal fluency (SVF) assessment. Speech was recorded through a mobile tablet built-in microphone and downsampled to 8 KHz to simulate telephone conditions. SVF sound segments were analyzed using Google’s ASR service for possible transcriptions. A variety of features were extracted from generated transcripts and evaluated by the SVM classifier. The overall error rate of the automatic transcripts was 33.4%, and the automated ASR classifier reached results comparable with those of the classifier trained on manual transcriptions. A. Lauraitis et al. [16] proposed neural impairment screening and self-assessment using mobile application for MCI detection based on the Self-Administered Gerocognitive Examination (SAGE) screening. They developed mobile application to collect data from different tasks. VF task was conducted by instructed participant to write down 12 different items in the given category as text field inside mobile application for SAGE score calculation. Voice recording performed as additional task to evaluate speech impairment by extracted several features for instance pitch, Mel-frequency cepstral coefficients (MFCC), Gammatone cepstral coefficients (GTCC) and spectral skewness for further speech analysis with SVM and bidirectional long short-term memory (BiLSTM) classifier with 100% and 94.29% accuracy respectively.

In our study, the ASR system plays an important role to enable reliable word detection for the MoCA language fluency test scoring system. Various approaches have been proposed for ASR with Thai language support over the years. For instance, Chaiwongsai et al. [17] proposed HMM-based isolated word speech recognition with a tone detection function to improve the accuracy for tonal language. The tone detection function was added as a parallel computation process to detect tone level and map with the results from speech recognition part to obtain the final results. Experimental results revealed that the accuracy of Thai word detection improved by 4.94% for TV remote control commands and by 10.75% for Thai words that had different meanings with each tone.

One approach to avoid new training steps whenever new words are added into the dictionary is the phoneme recognition approach. To this end, Theera-Umpon et al. [18] proposed a new method to classify the tonal accents of syllables using soft phoneme segmentation techniques for Thai speech, which was better than the hard-threshold approach for phoneme classification. Hu et al. [19] conducted an experiment with Mandarin and Thai by incorporating tonal information from fundamental frequency (F0) and fundamental frequency variation (FFV) into Convolution Neural Network (CNN) architecture and compared CNN with DNN. The WER for Thai was 33.19% and 35.16% for CNN and DNN, respectively.

Background

MoCA Scoring Criteria

The standard version of MoCA test is a one-page 30-point test available at [20]. Details on the specific MoCA items are as follows [8]. Visuospatial and executive functions are assessed using clock-drawing task (3 points), copy drawing task (1 point), trail making task (1 point), a two-item verbal abstraction task (2 points), confrontation naming task with low-familiarity animals (3 points). Language is assessed with repetition of complex sentences (2 points) and verbal fluency task (1 point). Short-term memory is evaluated by learning of nouns and delayed recall after 5 minutes (5 points). Attention and concentration are assessed using tapping test (1 point), serial subtraction (3 points), digits forward and backward (2 points). Finally, orientation to time and place is evaluated (6 points).

Speech Production System

Human speech production system can be divided into three parts: the system below the larynx, the larynx and its surrounding structures, and the structures and the airways above the larynx [21] as illustrated in Fig. 1.

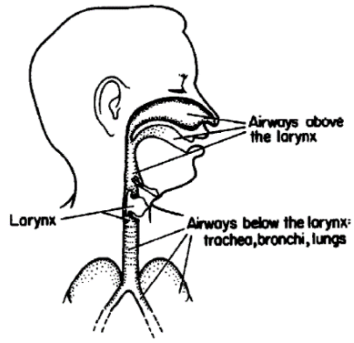


Figure 1. Representation of the three components of Speed Production System [21]

Source-Filter Model

Source-Filter model is a way of explaining acoustic of sound by modelling how the pulse produced by the glottis (source) are shaped by the vocal tract (filter) [22]. The airflow from the lungs which is modulated by the glottis acts as source signal passing through the vocal tract acts as a kind of filter or amplifier so changing the shape of vocal tract cavity by placing the tongue and the other articulators in particular position can cause different frequencies to be amplified as shown in Fig. 2 and 3.

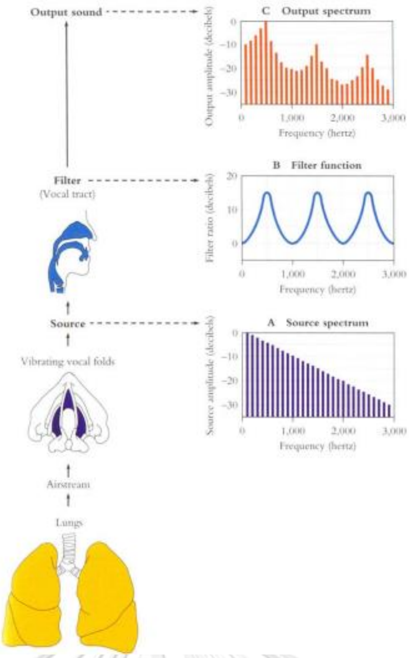


Figure 2. Source-Filter model (from http://www.xavieranguera.com/tdp_2011/4-Source-Filter-Models.pdf)

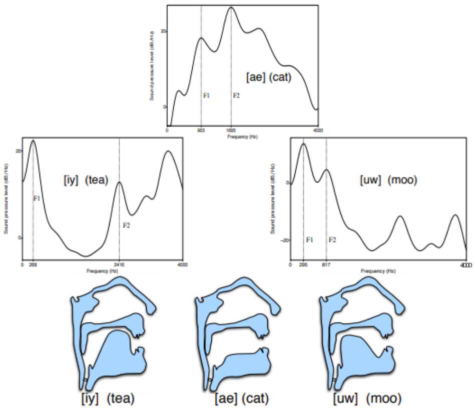


Figure 3. Visualizing the vocal tract position as a filter: the tongue positions for three English vowels [22]

Speech Recognition Architecture

A typical speech recognition system consists of basic components shown in Fig. 4. Acoustic front-end take care of converting speech waveform into appropriate feature vector with all necessary information for recognition using feature extraction techniques. The decoder uses both acoustic models, language models and pronunciation lexicon to generate words sequence that has maximum posterior probability for the input feature vectors.

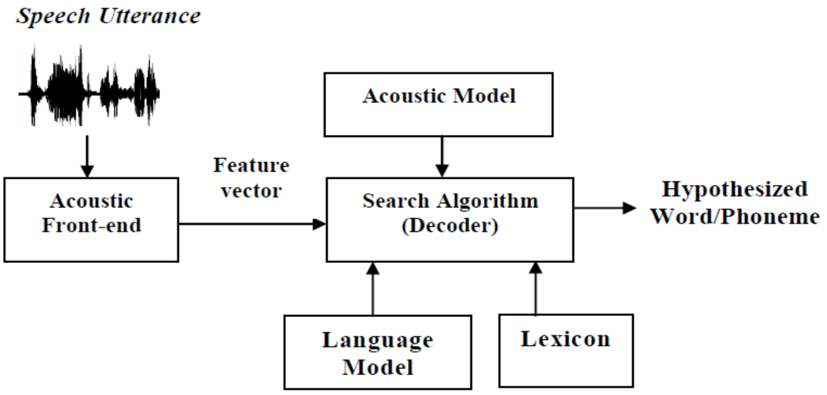


Figure 4. Speech Recognition Architecture [10]

Acoustic Front-end

Major task of acoustic front-end involve signal processing and feature extraction with the main goal to remove unwanted information, reduce dimensionality and extract important feature vectors from audio signal for further processing. The feature extraction normally performed in three stages [23]. The first stage performs spectra temporal analysis by convert signal from time-domain to frequency-domain and generates raw features describing the envelope of the power spectrum of short speech intervals. Second stage compiles an extended feature vector composed of statistic and dynamic features. The last stage transforms this extended feature vector into compact and robust vectors that can be supplied to recognizer.

Feature Extraction

There are various techniques to extract features but the most common in speech recognition is Mel-Frequency Cepstral Coefficient (MFCC) [22]. The MFCC feature extraction process has many steps which are shown in Fig. 5.

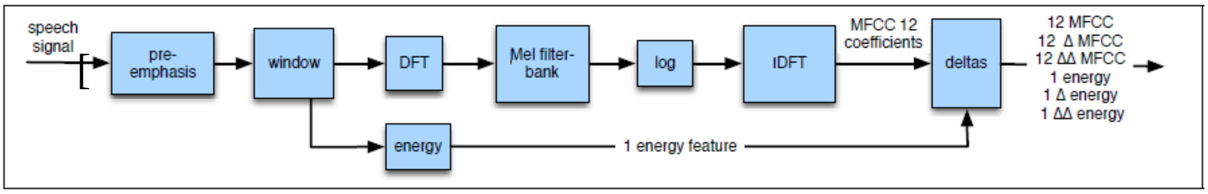


Figure 5. Extracting a sequence of 39-dimensional MFCC feature vectors from a quantized digitized waveform [22]

Step 1: Pre-emphasis

This step is used to boost the amount of energy in the high frequencies of input speech signal.

Step 2: Window

This step is used to slices the input signal into sliding frames. This is done by using a frame size of N milliseconds with frame shift of M milliseconds. A Hamming window is commonly used to avoid effect associate with discontinuities of signal at the boundaries caused by rectangular window.

Step 3: Discrete Fourier Transform (DFT)

DFT is applied to the windowed speech signal, resulting in the magnitude and phase representation of the signal at different frequency bands. A commonly used algorithm to compute DFT is the Fast Fourier Transform or FFT.

Step 4: Mel filter bank

This step is to warp the frequencies output from DFT onto the logarithmic Mel scale to simulate human hearing perception which is less sensitive at frequencies above

1000 Hz. A Mel frequency can be computed from the raw acoustic frequency as follows:

$$mel(f) \ = \ 1127ln\left(1+\frac{f}{700}\right) \tag{1}$$

A bank of filters known as triangular filter is comprised of 10 linearly-spaced below 1000 Hz and 10 log-spaced filters above 1000 Hz as shown in Fig. 6 to collect energy from each frequency band. The output of filtering the DFT signal by each Mel filter is known as the Mel spectrum.

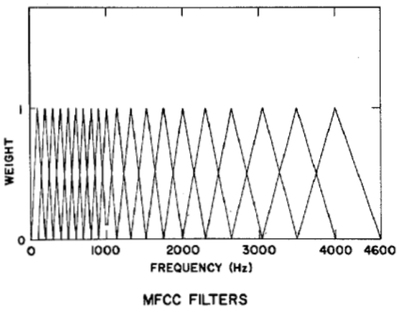


Figure 6. Filters for generating MFCC [24]

Step 5: Log

In this step, we take logarithm of Mel spectrum values as humans are less sensitive to small energy change at high energy than small changes at low energy level. In addition, using log makes feature less sensitive to variations in input.

Step 6: Inverse Discrete Fourier Transform (iDFT)

In this step, we need to separate the glottal source and the filter by computation of cepstrum using inverse DFT to convert log Mel spectrum into time domain. Result of conversion is called Mel Frequency Cepstral Coefficient where first 12 cepstral values will represent information about vocal tract filter which is useful for phone detection.

Step 7: Delta and energy

In addition to 12 cepstral coefficients, the energy from each frame is another feature that providing useful cue to identify phones. The energy in a frame for signal *x* in window from time sample *t1* to time sample *t2* is represented as:

$$Energy = \sum_{t=t1}^{t2} x^{2}[t] \qquad (2)$$

Speech signal is not constant from frame to frame so there is a need to add features related to change in cepstral features over time by adding each of 13 features (12 cepstral plus energy) a delta and double delta values end up with 39 MFCC features. Deltas is computed from difference between frames; delta value *d(t)* for cepstral value *c(t)* at time *t* can be estimated as:

$$\mathrm{d(t)} = \frac{c(t+1) - c(t-1)}{2} \qquad (3)$$

Acoustic Model

Acoustic modeling refers to process of establishing statistical representations of feature vector sequence to represent relationship between audio signal and linguistic units that make up speech which include knowledge about acoustics, phonetics, gender and dialect differences among speakers [25]. Hidden Markov Model (HMM) is one of the most used statistical models to build acoustic models for generating a sequence of phonemes as output. Acoustic model is created by taking a large database of speech called a speech corpus and using special training algorithms to create statistical representations for each phoneme in a language [23].

Language Model

Language models refer to collection of constraints on sequence of possible words that are likely to occur in a given language. These constraints can be represented by the rules of grammar or statistics of word pair estimated on a training corpus [23]. One of the most popular language model N-gram which predicts the occurrence of a word based on the occurrence of its N – 1 previous word. For example, in bigram (2-gram) language model the current word depends on last word only.

Materials and Methods

Data collection

Speech data were collected as part of the Digital MoCA project trial run in Chulalongkorn Hospital, Thailand, with 60 participants (52 women and 8 men) between the ages of 60 and 80. During the language fluency test, the patient was given one minute to orally produce as many Thai words as possible beginning with "ก." The whole utterance was recorded as single audio file through a standard built-in microphone on an iPad device with a sampling rate at 12 kHz in M4A file format stored in the Digital MoCA database, as shown in Fig. 7.

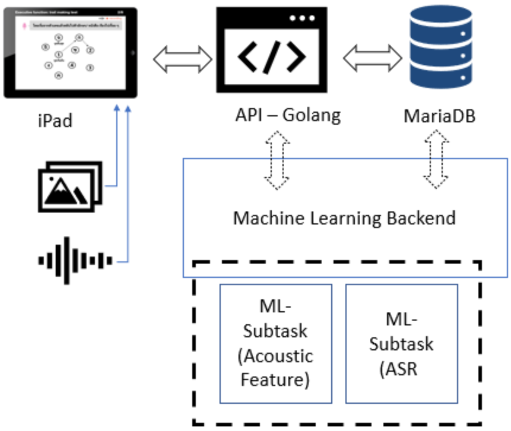


Figure 7. Digital MoCA system architecture overview

Data preprocessing

Several challenges became apparent in the speech data: for example, variation in the pronunciation of each patient, mixtures of conversation sentences with the intended word and background noise, and the sound of the physician’s digital pencil touching the iPad screen impacted the training dataset and accuracy of speech recognition. Several audio files that contained conversations between the patient and physician during the test were manually removed before further processing of the data.

To improve the data quality, we implemented preprocessing steps for the original speech data by applying a recurrent neural network noise suppression algorithm RNNoise [26] to remove the background noise from all utterances with an additional spectral gating filter to eliminate digital pencil sounds and generate clean speech data, as illustrated in Fig. 8.

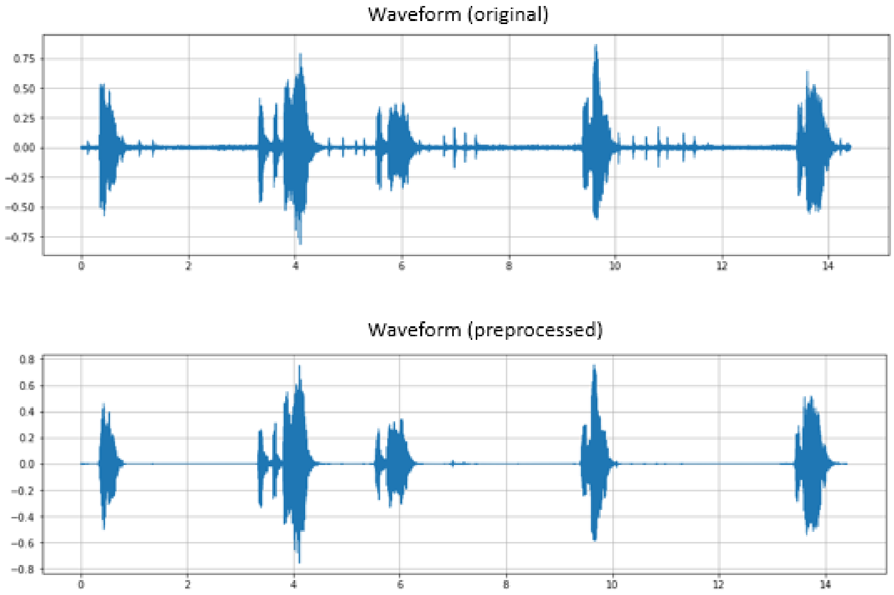


Figure 8. Patient’s speech waveform after preprocessing

Speech corpus and data augmentation

We did not have a sufficient amount of speech data for Thai words beginning with “ก” to develop a reliable ASR system on this task. We incorporated the large vocabulary Thai continuous speech recognition corpus called LOTUS [27] into the training data, where a phonetically distributed (PD) set was used for acoustic model training to cover most of the basic phone units in Thai. Data augmentation using frequency shifts at 100 Hz, 300 Hz, and 500 Hz was applied to both the original and clean speech data. Additional speed perturbation to generate more data with higher and lower speeds not more than 12% of the original speed was implemented to populate the high-resolution training dataset. The details of the training dataset are summarized in Table 1.

Table 1. Details of the training dataset

Corpus	Number of Utterances	Durations (Hrs.)
LOTUS – PD	3,040	5:24
Digital MoCA	552	1:16
Augmented Digital MoCA	1,518	4:39
Augmented LOTUS + MoCA	12,620	30:10

Model training

The Hidden Markov Model (HMM) was the main foundation for speech recognition for decades [28], and the Gaussian Mixture Model – Hidden Markov Model (GMM-HMM) system was a robust model used to develop the ASR system for many years. A novel hybrid model architecture called the Deep Neural Network - Hidden Markov Model (DNN-HMM) has been shown to outperform the GMM-HMM on a variety of speech recognition benchmarks [29] and has been widely used in speech recognition recently. However, achieving a good accuracy of DNN training normally requires a large amount of data, which is not available in our domain. To explore an optimal solution with the existing constraints, we developed an ASR system based on Kaldi [12] and created two training models using the standard GMM-HMM and DNN-HMM for comparison.

GMM-HMM acoustic model

Since Thai is a tonal language, we extracted MFCC together with pitch features to help improve the system's overall performance [30]. After these features were extracted, we then transformed with Cepstral Mean and Variance Normalization (CMVN) following the training sequence shown in Fig. 9.

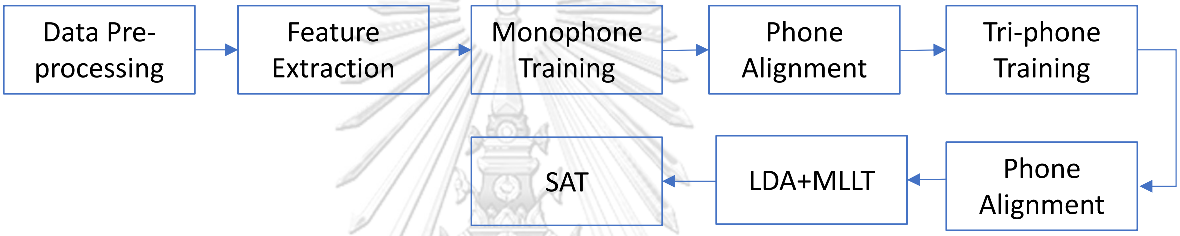


Figure 9. GMM-HMM model training sequence

In additional to the standard acoustic model mono-phone and tri-phone training, the Linear Discriminant Analysis (LDA) – Maximum Likelihood Linear Transform (MLLT) was applied to reduce feature vectors and derive unique transformation for each speaker, followed by normalization with Speaker Adaptive Training (SAT) to reduce inter-speaker variability in the training data [31].

DNN-HMM acoustic model

There are several types of DNN architecture supported by Kaldi, among which the most popular is TDNN [32]. It uses a lattice-free version of the MMI objective function called chain models, which use smaller frame rate at the output of the neural net to reduce computation at the test time, making the results faster to decode. The standard chain models training sequence relies on alignment data from the GMM-HMM stage as input to the TDNN layers depicted in Fig. 10.

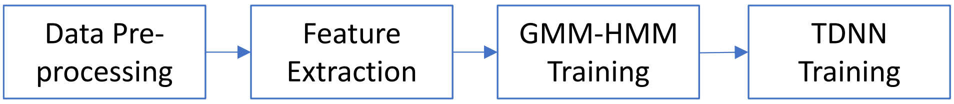


Figure 10. LF-MMI model training sequence

The architecture of the LF-MMI model comprises five TDNN layers with splicing indexes as (-1,0,1) (-1,0,1) (-3,0,3) (-3,0,3) (-3,0,3), as shown in Fig. 11. The (-1,0,1) means that the TDNN layer will process the input vectors at time t-1, t, and t+1, and (-3,0,3) means the TDNN layer will process input vectors at time t-3, t, and t+3.

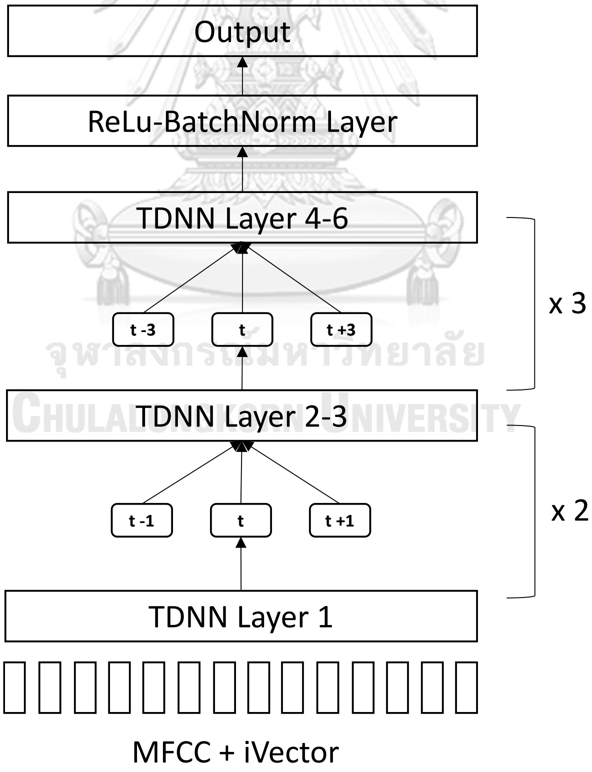


Figure 11. TDNN architecture for LF-MMI

Due to the limited amount of Digital MoCA training data, the quality of the GMM-HMM stage may not enable acceptable results with our chain models. We used another technique proposed by Hadian et al. [33] called end-to-end LF-MMI (EE-LF-MMI), where training was performed in a flat-start manner of single DNN in one stage without using any previously trained model or alignment data, as shown in Fig. 12.

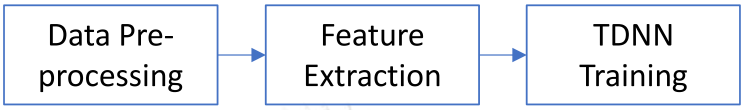


Figure 12. EE-LF-MMI model training sequence

The model architecture for end-to-end LF-MMI was composed of 13 TDNN layers with splicing indexes (−1,0,1) in layers 2–4 and (−3,0,3) in layers 6–13, as illustrated in Fig. 13.

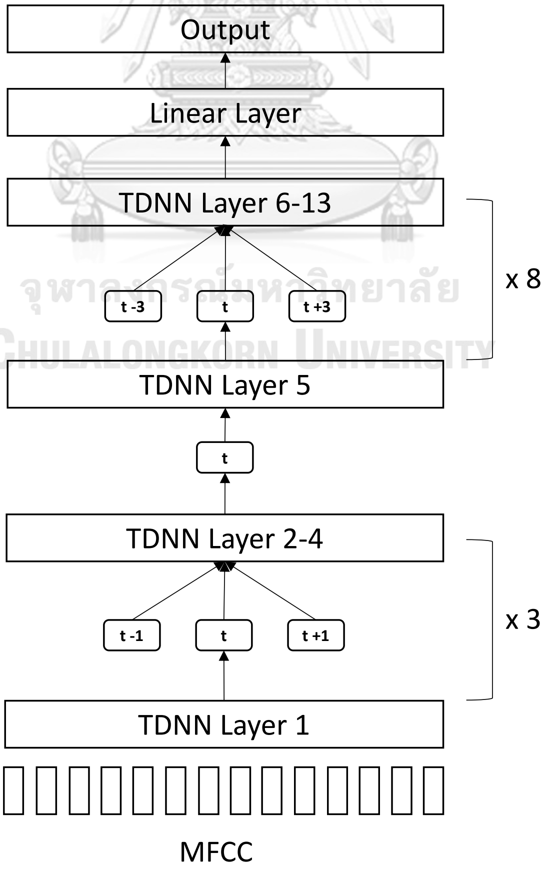


Figure 13. TDNN architecture for EE-LF-MMI

Language model and lexicon

A special dictionary was prepared by extracting all words starting with “ก” from a standard Thai dictionary, yielding 2,253 words. We combined all utterance transcriptions from the LOTUS corpus and Digital MoCA language fluency test dataset as well as the special dictionary to construct a text corpus for tri-gram language model training and lexicon creation.

Evaluation metrics

Performance of ASR systems is usually evaluated by the WER, which is calculated as follows in (4):

$$\mathrm{WER} = \frac{(I + S + D)}{N} \tag{4}$$

where, *I* is the number of insertions, *S* is the number of substitutions, *D* is the number of deletions, and *N* is the number of words in the reference.

To evaluate the performance of MoCA language fluency scoring, we defined a new metric called fluence score accuracy (FSA) to measure the difference between the final score rated manually by health professionals and the score calculated by the system. This is shown in (5) below:

$$FSA = \frac{TM}{(TM + TN)} \tag{5}$$

where, *TM* is the total number of utterances that received the same score in the manual and automatic calculations and *TN* is the total number of utterances that received different final scores in the manual and automatic calculations.

Decoding and Scoring

According to the scoring criteria of the MoCA test for language fluency, scores are given to words that start with “ก ” as defined in a standard Thai dictionary, excluding proper names and duplicate words. The final score is 1 if the total eligible word count is 11 or higher; otherwise, the final score is 0. Thus, the final score function can be calculated as shown in (6):

$$\text{Fluency score } \mathrm{f}(x) = \begin{cases} 0, & x < 11 \\ 1, & x \geq 11 \end{cases} \tag{6}$$

where $\mathbf{X}$ is the total number of unique words in the dictionary.

Since output word sequences might contain ineligible words caused by recognition process errors as well as unintended spoken words from patients, we implemented additional steps to filter words that begin with the initial phoneme /k/ corresponding to “ก ” in Thai before feed into the scoring algorithm. The steps for decoding and scoring are illustrated in Fig. 14.

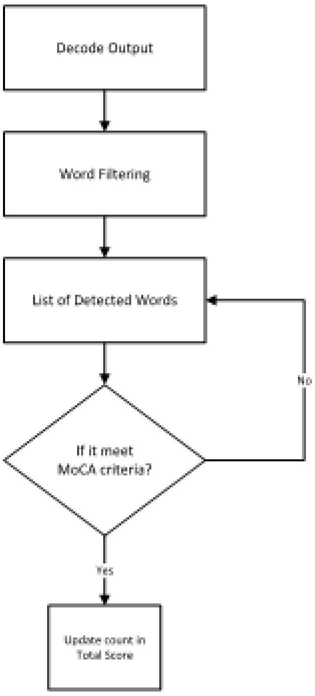


Figure 14. Decoding and scoring flowchart

Results and Discussion

We conducted an experiment with the original data from the Digital MoCA and LOTUS corpus using the GMM-HMM model as the baseline to confirm our hypothesis that incorporating data from the LOTUS corpus can improve the accuracy of our model. We performed GMM-HMM training with the same configuration applied to each dataset from the Digital MoCA, LOTUS, and a combined set from both corpora. The results showed that our GMM-HMM model achieved good accuracy for the LOTUS training samples but did not work with the data collected from the Digital MoCA due to the smaller sample size, while the combined dataset from LOTUS and the Digital MoCA improved the overall accuracy of the model when validated with data from the Digital MoCA, as shown in Table 2. We analyzed the results further and found that the main reason behind the poor results from the GMM-HMM model was a phone alignment problem that occurred during the training.

Table 2. Baseline ASR results

Dataset	Features	Model	Validation	WER
LOTUS	MFCC	GMM-HMM (triphone)	LOTUS	2.64%
LOTUS	MFCC + Pitch	GMM-HMM (triphone)	LOTUS	3.81%
Digital MoCA	MFCC	GMM-HMM (triphone)	MoCA	96.06%
Digital MoCA	MFCC + Pitch	GMM-HMM (triphone)	MoCA	91.34%
LOTUS	MFCC	GMM-HMM (triphone)	MoCA	100.39%
LOTUS	MFCC + Pitch	GMM-HMM (triphone)	MoCA	100.39%
LOTUS + MoCA	MFCC	GMM-HMM (triphone)	MoCA	74.41%
LOTUS + MoCA	MFCC + Pitch	GMM-HMM (triphone)	MoCA	84.65%

From the experiments, we observed that pitch features helped reduce WER in general, but no improvement was visible for the LOTUS dataset in the current setup. To evaluate the performance of our proposed model, we performed an experiment with the combined dataset with data augmentation using different configurations and features for GMM-HMM, TDNN-HMM (LF-MMI), and TDNN-HMM (EE-LF-MMI). The results showed that the proposed model EE-LF-MMI improved the overall accuracy of the recognizer without a need for alignment data from the previous training in the GMM-HMM model, as depicted in Table 3.

Table 3. Evaluation of ASR results

Dataset	Features	Model	WER
Augmented LOTUS + MoCA	MFCC	TDNN-HMM (EE-LF-MMI)	**41.30%**
Augmented LOTUS + MoCA	MFCC + iVector	TDNN-HMM (LF-MMI)	*
Augmented LOTUS + MoCA	MFCC + Pitch	GMM-HMM	93.48%
* Model failed to run due to bad alignment			

We used the phone alignment results from the GMM-HMM training with the LOTUS+MoCA dataset for the LF-MMI model, and the results confirmed that we could not obtain good improvement due to the bad quality of input. Decoded output from the EE-LF-MMI obtained a WER of 41.30%, which was relatively high, but it shown significant improvement over the other two models.

Currently, the regular LF-MMI achieves state-of-the-art results on several speech recognition tasks [33], while the end-to-end LF-MMI can obtain a comparable performance with the regular LF-MMI but with a simplified training pipeline and works well with a small dataset. In our study, the LF-MMI approach obtained better accuracy than the GMM-HMM, but the model architecture required alignment data from the previous training due to very poor results from the GMM-HMM. Thus, phone alignment information significantly affected the quality and accuracy of the TDNN layers. Meanwhile, the EE-LF-MMI approach demonstrated a high potential for future ASR development where domain-specific data are scarce with minimum effort needed for feature engineering.

Data augmentation analysis

We conducted a comparison to understand the impact of the data augmentation techniques used in our proposed model. We applied a combination of frequency shift and speed perturbation help to reduce errors and improve the overall accuracy, as shown in Table 4.

Table 4. Comparison of accuracy impact by data augmentation over combined datasets

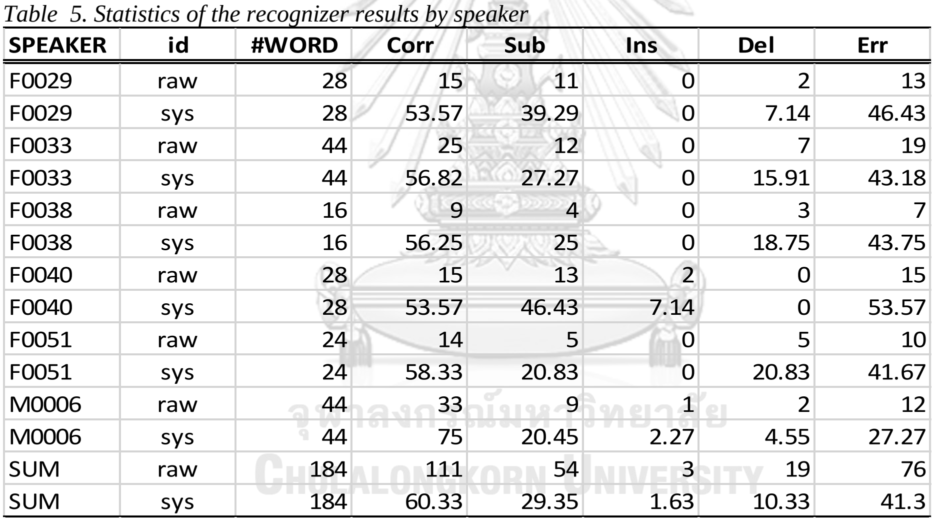
Model	Frequency shift	Speed purturbation	WER
TDNN-HMM (EE-LF-MMI)	N	Y	80.98%
TDNN-HMM (EE-LF-MMI)	Y	Y	**41.30%**

Word count and utterance analysis

We analyzed the results of the decoding output from our model for each speaker to see where the errors came from. Substitution error was the main contributor with 29.35% on average. Then, 1.63% of the problem came from insertion and 10.33% from deletion errors, while the percentage of words that could be detected correctly was in the range of 60.33%, as reported in Table 5.

We observed higher accuracy rates for male speaker compared to female speakers, so we listened to the original audio recordings and found that the voice quality of male speakers seemed to be better. As the audio samples contained both long utterances and single words, to gain a better understanding of the sequence of errors for long utterances, we filtered out the results from single words and analyzed only long utterances. By doing this, we found that major mismatches were caused by deletion and insertion errors, as shown in Table 6. Further analysis showed that the model could not accurately predict words with similar tones: e.g., ไกล /klay/ vs. ใกล้ /klây/, ไกล /klay/ vs. ไก่ /kày/, as shown in Fig. 15.

Table 5. Statistics of the recognizer results by speaker



SPEAKER	id	#WORD	Corr	Sub	Ins	Del	Err
F0029	raw	28	15	11	0	2	13
F0029	sys	28	53.57	39.29	0	7.14	46.43
F0033	raw	44	25	12	0	7	19
F0033	sys	44	56.82	27.27	0	15.91	43.18
F0038	raw	16	9	4	0	3	7
F0038	sys	16	56.25	25	0	18.75	43.75
F0040	raw	28	15	13	2	0	15
F0040	sys	28	53.57	46.43	7.14	0	53.57
F0051	raw	24	14	5	0	5	10
F0051	sys	24	58.33	20.83	0	20.83	41.67
M0006	raw	44	33	9	1	2	12
M0006	sys	44	75	20.45	2.27	4.55	27.27
SUM	raw	184	111	54	3	19	76
SUM	sys	184	60.33	29.35	1.63	10.33	41.3

Table 6. Detailed results of mismatches per utterance

Utterance		Decode output																					
F0029_1001	ref	การบ้าน	การงาน	ไก่	ก้าว	ก้าน	กน	กลม	กรุ่น	โกง	กิน	กับข้าว	ก้อน	เกลี่ย	กก								
F0029_1001	hyp	การบ้าน	การงาน	ไก่	ก้าว	ก้าน	กล	กลม	กรุ่น	โกง	***	เกม	กัด	กาม	กลม								
F0029_1001	op	C	C	C	C	C	S	C	C	C	D	S	S	S	S								
F0029_1001	#csid	8	5	0	1																		
F0033_1001	ref	กินข้าว	กลับบ้าน	กรุงเทพ	การบ้าน	กลับหลังหัน	กุ้ง	ไก่	กลืน	กรุบกริบ	กิ๊ก	กุ๊กกู๋	ก่อนนอน	กังหัน	กลม	กลิ้ง	กรอกหน้า	กลับหลังหัน	กลับไปกลับมา	กาญจนบุรี	ไกล	ใกล้	ก้อง
F0033_1001	hyp	***	***	***	การบ้าน	กลับหลังหัน	กุ้ง	ไก่	กลืน	กรุบกริบ	กิ๊ก	กุ๊กกู๋	ก่อนนอน	กังหัน	กลม	กลิ้ง	กรอกหน้า	กลับหลังหัน	กรรม	เกวียน	ไกล	ใกล้	ก้อง
F0033_1001	op	D	D	D	C	C	C	C	C	C	C	C	C	C	C	C	C	C	S	S	C	C	C
F0033_1001	#csid	17	2	0	3																		
F0038_1001	ref	เกิด	ใกล้	กิน	กาง	เกี้ยว	กอง	เก็บ	กิน														
F0038_1001	hyp	***	ไกล	กาย	กาง	เกี้ยว	กอง	เก็บ	กิน														
F0038_1001	op	D	S	S	C	C	C	C	C														
F0038_1001	#csid	5	2	0	1																		
F0040_1001	ref	แกง	ก้อน	กิน	***	กา	กบ	เกาะ	กาง	ไก่	กก	ใกล้	ไกล	***	ก้น	กอง	ก้าง						
F0040_1001	hyp	แกง	ก้อน	กิน	กาง	กด	ก้อน	เกาะ	กาง	ไก่	กก	ใกล้	ไกล	กล	ก่น	เกรง	กลอน						
F0040_1001	op	C	C	C	I	S	S	C	C	C	C	C	C	I	S	S	S						
F0040_1001	#csid	9	5	2	0																		
F0051_1001	ref	ไก่	กิ้งก่า	กา	กิ่ง	ก้อน	กิน	กิ้งกือ	กง	เกวียน	ไกล	ใกล้	กุ้ง										
F0051_1001	hyp	ไก่	กิ้งก่า	กา	กิ่ง	ก้อน	กิน	***	กง	เกวียน	ไกล	***	กาย										
F0051_1001	op	C	C	C	C	C	C	D	C	C	C	D	S										
F0051_1001	#csid	9	1	0	2																		
M0006_1001	ref	กระดาน	กระดาษ	กระเด้ง	กระดอน	กระเป๋า	กระปุก	กะปิ	กา	กระติก	กติกา	ก่อเกิด	เกิด	กิจการ	กรรม	กีฬา	กิเลส	กำบัง	กังหัน	กังวล	กังวาน	กึกก้อง	กระดุม
M0006_1001	hyp	กระดาน	กระดาษ	กระเด้ง	กระดอน	กระเป๋า	กระปุก	กะปิ	กา	กระติก	กติกา	ก่อเกิด	เกิด	กิจการ	กรรม	กีฬา	กิเลส	กำบัง	กังหัน	กังวล	กังวาน	กึกก้อง	กระดุม
M0006_1001	op	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
M0006_1001	#csid	22	0	0	0																		

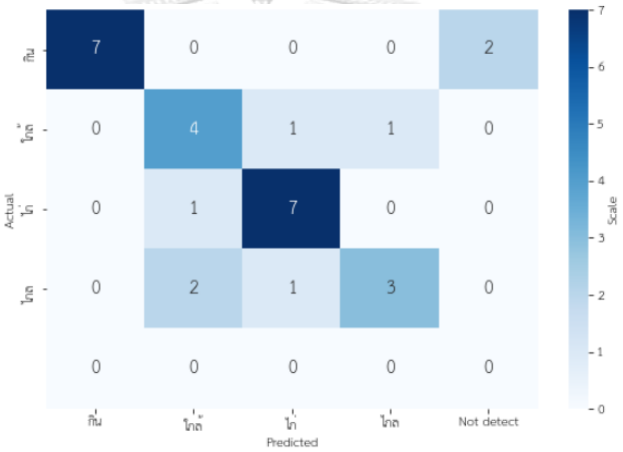


Figure 15. Confusion matrix for words with similar tone

We conducted further analysis of the distribution of words in the MoCA training set to find the most common words uttered by patients (see Fig. 16). We compared the top 10 words from the test data with the total distribution in the training set. The results confirmed that the model could predict correct words outside the training data, as illustrated in Table 7.

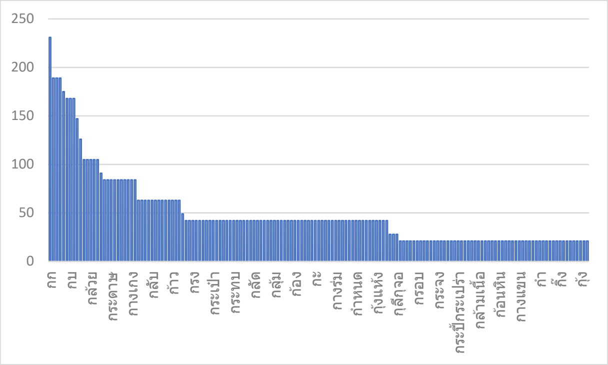


Figure 16. Word count from the MoCA training data

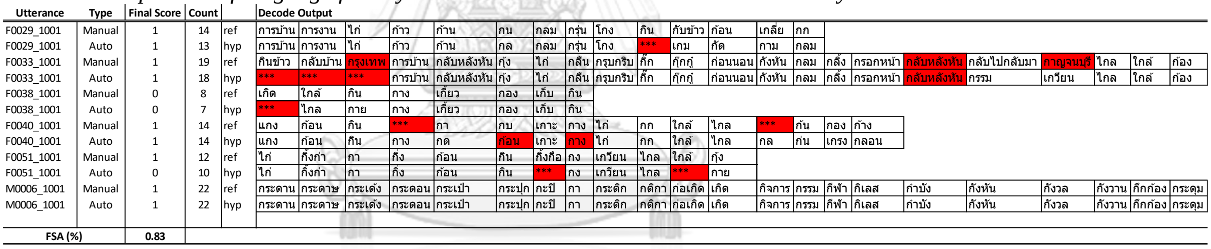
Table 7. Top 10 words from test utterances

Word	Count	Word count in training	Correctly detected
กิน	10	175	7
ใกล้	8	0	4
ไก่	8	0	7
ไกล	6	0	3
ก้อน	6	84	5
กา	6	126	4
เกิด	4	0	2
กก	4	0	3
กลม	4	0	3
กลับหลังหัน	4	0	3

In this study, we have developed and evaluated a new approach to integrate ASR techniques into a MoCA assessment tool for conducting a language fluency test with Thai language support. Most previous studies focused on utilizing ASR to assess verbal semantic fluency tasks in English, for which ample speech data are publicly available. This novel method is proposed to utilize ASR for assisting a phonemic fluency task in Thai, which is the first attempt of its kind. Several challenges still need to be addressed, starting from the data collection process, which directly affected the quality of voice recordings and had major impacts on the overall accuracy of ASR. Acoustic model training with low resources, namely, no domain-specific data available, was one of the biggest challenges in the speech recognition task.

To evaluate the usability of an automated scoring system for the MoCA assessment tool, we examined the results of automated scoring compared with manual scoring by health professionals after the assessment. For those utterances with substitution errors, the ASR recognizer tried to predict the closest words that were still within the dictionary results with no big deviation from the total count of words and no impact to the overall score for that utterance. The final scoring calculated by the system showed an accuracy of 83%, as reported in Table 8, indicating a high potential for further development and use in clinical practice.

Table 8. Comparison of language fluency scores between the manual and automated systems



Utterance	Type	Final Score	Count		Decode Output																					
F0029_1001	Manual	1	14	ref	การบ้าน	การงาน	ไก่	ก้าว	ก้าน	กน	กลม	กรุ่น	โกง	กิน	กับข้าว	ก้อน	เกลี่ย	กก								
F0029_1001	Auto	1	13	hyp	การบ้าน	การงาน	ไก่	ก้าว	ก้าน	กล	กลม	กรุ่น	โกง	***	เกม	กัด	กาม	กลม								
F0033_1001	Manual	1	19	ref	กินข้าว	กลับบ้าน	กรุงเทพ	การบ้าน	กลับหลังหัน	กุ้ง	ไก่	กลืน	กรุบกริบ	กิ๊ก	กุ๊กกู๋	ก่อนนอน	กังหัน	กลม	กลิ้ง	กรอกหน้า	กลับหลังหัน	กลับไปกลับมา	กาญจนบุรี	ไกล	ใกล้	ก้อง
F0033_1001	Auto	1	18	hyp	***	***	***	การบ้าน	กลับหลังหัน	กุ้ง	ไก่	กลืน	กรุบกริบ	กิ๊ก	กุ๊กกู๋	ก่อนนอน	กังหัน	กลม	กลิ้ง	กรอกหน้า	กลับหลังหัน	กรรม	เกวียน	ไกล	ใกล้	ก้อง
F0038_1001	Manual	0	8	ref	เกิด	ใกล้	กิน	กาง	เกี้ยว	กอง	เก็บ	กิน														
F0038_1001	Auto	0	7	hyp	***	ไกล	กาย	กาง	เกี้ยว	กอง	เก็บ	กิน														
F0040_1001	Manual	1	14	ref	แกง	ก้อน	กิน	***	กา	กบ	เกาะ	กาง	ไก่	กก	ใกล้	ไกล	***	กัน	กอง	ก้าง						
F0040_1001	Auto	1	14	hyp	แกง	ก้อน	กิน	กาง	กด	ก้อน	เกาะ	กาง	ไก่	กก	ใกล้	ไกล	กล	ก่น	เกรง	กลอน						
F0051_1001	Manual	1	12	ref	ไก่	กิ้งก่า	กา	กิ่ง	ก้อน	กิน	กิ้งกือ	กง	เกวียน	ไกล	ใกล้	กุ้ง										
F0051_1001	Auto	0	10	hyp	ไก่	กิ้งก่า	กา	กิ่ง	ก้อน	กิน	***	กง	เกวียน	ไกล	***	กาย										
M0006_1001	Manual	1	22	ref	กระดาน	กระดาษ	กระเด้ง	กระดอน	กระเป๋า	กระปุก	กะปิ	กา	กระติก	กติกา	ก่อเกิด	เกิด	กิจการ	กรรม	กีฬา	กิเลส	กำบัง	กังหัน	กังวล	กังวาน	กึกก้อง	กระดุม
M0006_1001	Auto	1	22	hyp	กระดาน	กระดาษ	กระเด้ง	กระดอน	กระเป๋า	กระปุก	กะปิ	กา	กระติก	กติกา	ก่อเกิด	เกิด	กิจการ	กรรม	กีฬา	กิเลส	กำบัง	กังหัน	กังวล	กังวาน	กึกก้อง	กระดุม
FSA (%)		0.83																								

Conclusion

The MoCA is a widely used tool of assessment for MCI detection, but data analysis options for the current paper-and-pencil based version are limited and require great manual efforts in data collection. Therefore, we seek to develop the Digital MoCA in Thailand. This research paper is a sub-project of the larger Digital MoCA project with the aim of creating technology to support automatic data collection and analysis to enable physicians and other health professionals in Thailand to conduct reliable cognitive assessments for a broader range of patients.

This paper demonstrated the possibilities to utilizing ASR techniques for word detection to assist with the MoCA language fluency test scoring system for Thai. The proposed method yields acceptable accuracy under a number of constraints where domain-specific data are not publicly available. However, more data need to be collected for the Digital MoCA, which is still in the early phase of development. In addition, several other challenges emerged in this study, mainly arising from data quality issues.

We see great potential to further improve the accuracy of the system, such as by enhancing the acoustic model with more data from patients, increasing the accuracy of the Thai tone detection with a better algorithm, and integrating a dedicated microphone system into the iPad device to differentiate the voices of patients from health professionals, which will help to improve the quality of the voice recording. In additional to benefits within the medical domain, the techniques developed from this research, such as isolated word recognition of Thai words beginning with “ก,” can be used as a baseline for future expansion of complex ASR system integration in various speech recognition domains.

REFERENCES

[1] I. J. Deary *et al.*, “Age-associated cognitive decline,” *Br. Med. Bull.*, vol. 92, no. 1, pp. 135–152, 2009, doi: 10.1093/bmb/ldp033.

[2] R. Y. Lo, “The borderland between normal aging and dementia,” *Tzu Chi Med. J.*, vol. 29, no. 2, pp. 65–71, 2017, doi: 10.4103/tcmj.tcmj_18_17.

[3] S. A. Gale, D. Acar, and K. R. Daffner, “Dementia,” *Am. J. Med.*, vol. 131, no. 10, pp. 1161–1169, 2018, doi: 10.1016/j.amjmed.2018.01.022.

[4] S. Duong, T. Patel, and F. Chang, “Dementia: What pharmacists need to know,” *Can. Pharm. J.*, vol. 150, no. 2, pp. 118–129, 2017, doi: 10.1177/1715163517690745.

[5] Y. E. Geda, “Mild cognitive impairment in older adults,” *Curr. Psychiatry Rep.*, vol. 14, no. 4, pp. 320–327, 2012, doi: 10.1007/s11920-012-0291-x.

[6] R. C. Petersen *et al.*, “Current Concepts in Mild Cognitive Impairment,” 2001. [Online]. Available: https://jamanetwork.com/.

[7] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “‘Mini-mental state’: A practical method for grading the cognitive state of patients for the clinician,” *J. Psychiatr. Res.*, vol. 12, no. 3, pp. 189–198, 1975, doi: https://doi.org/10.1016/0022-3956(75)90026-6.

[8] Z. S. Nasreddine *et al.*, “The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment,” *J. Am. Geriatr. Soc.*, vol. 53, no. 4, pp. 695–699, 2005, doi: 10.1111/j.1532-5415.2005.53221.x.

[9] S. Ahmed, C. de Jager, and G. Wilcock, “A comparison of screening tools for the

assessment of Mild Cognitive Impairment: Preliminary findings," *Neurocase*, vol. 18, no. 4, pp. 336–351, Aug. 2012, doi: 10.1080/13554794.2011.608365.

[10] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 1, no. 1, pp. 112–124, 2015, doi: https://doi.org/10.1016/j.dadm.2014.11.012.

[11] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in Alzheimer's disease and in its assessment," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 1948–1952, 2016, doi: 10.21437/Interspeech.2016-1228.

[12] D. Povey, G. Boulianne, L. Burget, P. Motlicek, and P. Schwarz, "The Kaldi Speech Recognition," *IEEE 2011 Work. Autom. Speech Recognit. Underst.*, no. January, 2011, [Online]. Available: http://kaldi.sf.net/.

[13] Z. Shao, E. Janse, K. Visser, and A. S. Meyer, "What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults," *Front. Psychol.*, vol. 5, no. JUL, pp. 1–10, 2014, doi: 10.3389/fpsyg.2014.00772.

[14] S. V. S. Pakhomov, S. E. Marino, S. Banks, and C. Bernick, "Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency," *Speech Commun.*, vol. 75, pp. 14–26, 2015, doi: 10.1016/j.specom.2015.09.010.

[15] J. Tröger, N. Linz, A. König, P. Robert, and J. Alexandersson, "Telephone-based dementia screening i: Automated semantic verbal fluency assessment," *ACM Int. Conf. Proceeding Ser.*, pp. 59–66, 2018, doi: 10.1145/3240925.3240943.

[16] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, “A Mobile Application for Smart Computer-Aided Self-Administered Testing of Cognition, Speech, and Motor Impairment,” *Sensors*, vol. 20, no. 11. 2020, doi: 10.3390/s20113236.

[17] J. Chaiwongsai, W. Chiracharit, K. Chamnongthai, and Y. Miyanaga, “An architecture of HMM-based isolated-word speech recognition with tone detection function,” *2008 Int. Symp. Intell. Signal Process. Commun. Syst. ISPACS 2008*, pp. 11–14, 2009, doi: 10.1109/ISPACS.2009.4806710.

[18] N. Theera-Umpon, S. Chansareewittaya, and S. Auephanwiriyakul, “Phoneme and tonal accent recognition for Thai speech,” *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13254–13259, Sep. 2011, doi: 10.1016/j.eswa.2011.04.142.

[19] X. Hu, M. Saiko, and C. Hori, “Incorporating tone features to convolutional neural network to improve Mandarin/Thai speech recognition,” in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 2014, pp. 1–5, doi: 10.1109/APSIPA.2014.7041576.

[20] “MoCA website. www.mocatest.org.” .

[21] K. Stevens, “Acoustic Phonetics,” in *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, vol. 109, 2000, p. 607.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2Nd Edition).* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009.

[23] K. S and E. Chandra, “A Review on Automatic Speech Recognition Architecture and Approaches,” *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 9, pp. 393–404, Apr. 2016, doi: 10.14257/ijsip.2016.9.4.34.

[24] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, 1980, doi: 10.1109/TASSP.1980.1163420.

[25] X. Huang and L. Deng, “An overview of modern speech recognition,” in *Handbook of Natural Language Processing, Second Edition*, 2010, pp. 339–366.

[26] “RNNoise - Recurrent neural network for audio. https://github.com/xiph/rnnoise.” .

[27] S. Kasuriya, V. Sornlertlamvanich, and P. Cotsomrong, “Thai Speech Corpus for Speech Recognition,” no. January 2003, 2014.

[28] B. Juang and L. Rabiner, “Automatic Speech Recognition - A Brief History of the Technology Development,” Jan. 2005.

[29] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.

[30] P. Ghahremani, B. Babaali, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. May, pp. 2494–2498, 2014, doi: 10.1109/ICASSP.2014.6854049.

[31] J. Mcdonough and J. Makhoul, “Speaker adaptive training: a maximum likelihood approach to speaker normalization,” pp. 6–9, 1997.

[32] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, “Phoneme Recognition Using Time-Delay Neural Networks,” *IEEE Trans. Acoust.*, vol. 37, no. 3, pp. 328–339, 1989, doi: 10.1109/29.21701.

[33] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, “End-to-end speech recognition using lattice-free MMI,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. 1, pp. 12–16, 2018, doi: 10.21437/Interspeech.2018-1423.

VITA

NAME	Pimarn Kantithammakorn
DATE OF BIRTH	24 March 1972
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Department of Physics Faculty of Science Kasetsart University
HOME ADDRESS	99/11 Ratchapreuk Rd., Om Kret Pak Kret Nonthaburi, 11120 Thailand